Michel Huguier et Pierre-Yves Boëlle

Biostatistiques pour le clinicien



Biostatistiques pour le clinicien

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Biostatistiques pour le clinicien

Michel Huguier et Pierre-Yves Boëlle



Michel Huguier
Service de chirurgie digestive
Hôpital Tenon
4, rue de la Chine
75970 Paris Cedex 20

Pierre-Yves Boëlle
INSERM U 707
Hôpital Saint-Antoine
184, rue du Faubourg-Saint-Antoine
75571 PARIS CEDEX 12

ISBN 978-2-8178-0463-7 Springer Paris Berlin Heidelberg New York



Maquette de couverture : Jean-François Montmarché
Mise en page : Desk

Sommaire

Troisième partie
Forces d'association, études multifactorielles, mesures d'impact, causalité

Sixième partie
Les évaluations pronostiques

Septième partie
Epidémiologie

Les auteurs

Michel Huguier est professeur honoraire de chirurgie digestive. Il s'est initié aux biostatistiques pour mener à bien des travaux de recherche clinique avec la collaboration de biostatisticiens, François Grémy, Claude Chastang, Jean-Claude Manderscheid, Antoine Flahault. Il a fait des enseignements de biostatistiques pour des cliniciens à Beyrouth, Bucarest, Hanoi, Montevideo, Paris, Saigon, Strasbourg, Toulouse, Tours.
Il est auteur de 240 publications dans des revues avec comités de lecture dont 80 dans des périodiques internationaux anglo-saxons.

Pierre-Yves Boëlle est ingénieur civil des Mines, professeur de biostatistiques à l'université Paris 6. Il enseigne les biostatistiques à la faculté de médecine, depuis la première année jusqu'au master. Il a collaboré avec de nombreux cliniciens pour l'analyse d'études dans des domaines aussi variés que la réanimation, l'anatomopathologie, l'orthopédie, l'oncologie, les maladies infectieuses, etc. Il est également chercheur dans une unité INSERM spécialisée dans la surveillance et la modélisation des maladies transmissibles.
Il est auteur de 120 publications dans des revues avec comités de lecture.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Introduction

La plupart des ouvrages de biostatistiques ont un abord très mathématique. Des remarques sur la toile montrent qu'ils ne sont pas toujours aisément accessibles, même à des étudiants en médecine dont la quasi-totalité vient de passer un baccalauréat scientifique. De plus, leur finalité médicale n'apparaît pas toujours clairement. Le présent livre a pour but de pallier, au moins en partie, à cette double constatation. Son originalité est d'avoir été écrit par un clinicien en collaboration avec un biostatisticien.

Un ouvrage indispensable. Pourquoi ?

Les progrès de la médecine sont le fruit d'innovations. Cependant, les innovations ne font pas toujours progresser l'élaboration d'un diagnostic ou bien l'efficacité et la tolérance d'un traitement ou encore la connaissance des facteurs de risque d'apparition d'une maladie ou d'un pronostic. L'histoire de la médecine montre que ce qui avait paru être un progrès n'a pas toujours été confirmé.

L'évaluation des innovations est indispensable. Elle seule évite ou réduit le temps pendant lequel on s'engage sur de fausses pistes, c'est-à-dire où l'on croit faire bénéficier les malades d'un progrès médical, alors qu'il n'en est rien. De plus, ces errements sont de plus en plus coûteux, notamment pour la solidarité nationale qui prend en charge les dépenses individuelles de soins.

La connaissance des méthodes d'évaluation est indispensable pour les auteurs d'un travail afin de le réaliser avec un maximum de rigueur méthodologique. Mais elle permet aussi aux lecteurs de se faire une opinion plus critique, plus scientifique sur les publications qui les submergent ou les sollicitations dont ils font l'objet. Pour les étudiants en médecine, c'est bien l'objectif d'une des épreuves de l'examen classant national à la fin du deuxième cycle des études.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Un ouvrage accessible à tous. Comment ?

Nous avons voulu que ce livre soit accessible à tout lecteur, même s'il n'a pas suivi une classe préparatoire de mathématiques supérieures ou s'il a, en partie, oublié ce qui avait pu lui être enseigné au lycée.
Pour ce faire, nous avons choisi de traiter la méthodologie de façon plus explicative que mathématique ; la compréhension des concepts nous a paru plus importante que la connaissance des démonstrations mathématiques sur lesquelles elle s'appuie.

Partie 1

Les données fondamentales Les différentes variables et leur mesure

Introduction

Tout travail scientifique, qu'il soit expérimental ou clinique, doit reposer sur quatre définitions clairement préétablies qui définissent le protocole de l'étude (ou plan expérimental lorsque l'on est en situation d'expérience). Un de ses buts est d'assurer la possibilité pour d'autres groupes d'investigateurs de reproduire le travail qui a été réalisé.
Il répond aux quatre questions suivantes (tableau I) :
– sur quoi a-t-on travaillé ?
– comment a-t-on travaillé ?
– qu'a-t-on cherché à évaluer ?
– quels ont été les critères de jugements de cette évaluation (ainsi que la manière dont ils ont été analysés, c'est-à-dire les méthodes statistiques utilisées).

Tableau I – Les quatre définitions fondamentales.

Sur quoi a-t-on travaillé ?
Le matériel d'étude (par exemple des souris, des hommes).
Comment a-t-on travaillé ?
Méthode de travail (par exemple comment les données ont été recueillies, prospectivement ou rétrospectivement).
Ce que l'on a cherché à évaluer ?
Un examen biologique avec la définition de sa normalité, un traitement avec la définition de sa posologie, de son mode d'administration, un facteur de risque, etc.
Quels ont été le(s) critère(s) de jugement ?
Maladie ou absence de maladie, efficacité et toxicité d'un médicament, récidive, survie, etc. ainsi que la façon dont ils ont été analysés (méthodes statistiques).

En corollaire, pour un lecteur, le contrôle de la qualité de ces définitions est un élément essentiel de la lecture critique et de l'interprétation des résultats. Ce contrôle est aisé. Il se fait en lisant la section « Matériel et méthodes » de l'article. Il s'agit de vérifier que les quatre définitions fondamentales ont bien été données, de façon précise. Il

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

est aisé, même pour un lecteur un peu entraîné, de reconnaître facilement les travaux dans lesquels ces définitions sont précises et claires. Dans le cas contraire, il s'agit de travaux qui sont habituellement mal conçus dès le départ. De ce fait, ils n'ont guère de chances d'apporter des informations utiles. Il est alors conseillé, sans grand risque, d'en arrêter la lecture.

Ensuite, « dans les sciences expérimentales, la mesure des phénomènes est un point fondamental », écrivait Claude Bernard [1]. De façon générale, l'évaluation biologique doit être aussi précise que possible. Elle se fait par l'appréciation de variables dont les valeurs observées dépendent de l'échantillon que l'on a constitué. Pour cette raison, ces variables sont dites aléatoires. Les variables se différencient des constantes dont une des plus connues est le nombre $\pi = 3{,}14116\ldots$ pour calculer la circonférence d'un cercle à partir de la valeur de son rayon. Rappelons aussi, en art, le nombre d'or Φ qui est égal à 1,618. C'est un rapport entre largeur et hauteur, baptisé « divine proportion » que l'on trouve dans des temples grecs anciens, ou dans le dessein de l'homme de Vitruve par Léonard de Vinci, celui qui est inscrit dans un cercle et un carré.

On peut distinguer les variables selon des caractéristiques qui ne sont pas exclusives, mais qui appelleront des traitements ou des interprétations appropriées (tableau II).

Tableau II – Les caractéristiques des variables.

Les variables qualitatives et quantitatives
– *Les variables qualitatives (ou catégorielles)* sont des variables qui sont appréciées selon qu'elles sont présentes ou absentes, par exemple, l'existence ou non d'une récidive dans une maladie ; ou qui correspondent à une caractéristique non quantitative de l'individu, par exemple, le département de résidence ou le pays de naissance.
– *Les variables quantitatives (ou numériques)* sont des variables dont les valeurs sont appréciées sous une forme numérique ; par exemple, la taille en centimètres, le poids en grammes, la glycémie en millimoles.

Les variables objectives et subjectives
– *Les variables objectives* sont mesurables directement comme les variables qualitatives et quantitatives.
– *Les variables subjectives* n'ont pas de référentiel absolu partagé pour toutes les observations. Ce sont, par exemple, une douleur ou encore la qualité de vie.

Les variables observées et censurées
– *Les variables observées* sont celles dont la valeur est connue par l'observation.
– *Les variables censurées* sont des variables pour lesquelles on n'observe pas exactement la valeur. Ce sont surtout des variables dont l'observation renvoie au temps, par exemple, la survie ou la survenue d'une récidive de maladie. Mais cela peut être également des dosages lorsque la valeur est inférieure au seuil de détection.

La description que l'on fait d'une variable résulte d'un choix. Il est parfois possible de modifier les caractéristiques d'une variable.
1. Une variable quantitative peut être transformée en variable qualitative, en choisissant une (ou des) valeur(s) seuil qui définira des classes ; ainsi, une variable quantitative, comme un amaigrissement, peut être transformée en variable qualitative à deux classes : amaigrissement de moins de 4 kg, ou de plus de 4 kg. De même, une variable censurée, comme la survie, peut, dans certaines conditions, être transformée en variable qualitative : survie à cinq ans ou non. Ces transformations font néanmoins perdre de l'information.
2. Une variable subjective peut être transformée en variable objective si l'on parvient à trouver un référentiel commun aux observations. Il existe pour cela plusieurs méthodes que nous indiquerons.
3. D'autres transformations peuvent être souhaitables ou habituelles. Dans l'infection à VIH, par exemple, on présente souvent les charges virales (nombre de copies de virions/mL) en logarithme à base 10, c'est-à-dire qu'une charge virale de 1 million (= 10^6) est représenté par 6 sur l'échelle log. L'utilisation de telles transformations permet de modifier la distribution, par exemple pour la rendre plus proche de la loi normale (de Laplace-Gauss).
Dernière notion : des **variables** sont dites **dépendantes** si leurs valeurs changent conjointement, par exemple, les valeurs du cholestérol total et du cholestérol estérifié. Dans le cas contraire, on parle de **variables indépendantes** comme la numération des hématies d'une part, et le dosage des phosphatases alcalines dans le sang d'autre part.

Référence

1. Bernard C (1865) Introduction à la médecine expérimentale. Baillière, Paris, p. 226

Les données fondamentales 1

Tout travail doit être élaboré avec un objectif précis, défini dans l'introduction du compte rendu de la recherche. Ensuite, comme nous l'avons indiqué, le chapitre « Matériel et méthodes » doit comprendre quatre descriptions fondamentales :
– ce sur quoi on a travaillé ;
– quelle a été la méthode de travail ;
– ce que l'on a cherché à évaluer ;
– quels ont été les critères de jugement de cette évaluation.
Ces données sont fondamentales pour permettre, soit de reproduire l'étude, soit pour chercher à expliquer des différences de résultats avec ceux d'une autre étude qui avait un objectif similaire ou assez proche. Pour un lecteur, les principes de la lecture critique ne sont que le corollaire des mêmes principes de l'élaboration d'un travail scientifique.

Le matériel d'étude

Le matériel d'une étude est constitué par ce *sur qui* (personnes, animaux, bactéries, virus, etc.) ou ce *sur quoi* (prélèvement tissulaire, sérum, urines, etc.) le travail a porté.

Des personnes

Qu'il s'agisse d'une étude clinique ou épidémiologique, il convient de préciser deux données : l'une concerne les critères d'inclusion des sujets dans l'étude, l'autre, la description de la population étudiée.
Les critères qui ont permis l'inclusion de chaque sujet dans l'étude définissent les caractéristiques de l'échantillon. Il est parfois nécessaire, surtout s'ils ne sont pas symétriques, de définir des critères d'exclusion de l'étude. Il est évident que la portée des résultats ne pourra concerner

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

qu'une population similaire à celle qui a été définie par les critères d'inclusion. En épidémiologie, comme nous le verrons, ces notions sont particulièrement importantes dans les *enquêtes cas-témoins* ou dans les *études de cohortes* qui doivent être le plus représentatives possible des populations étudiées.

Citons quelques exemples de critères d'inclusion ou d'exclusion courants. Dans des études cliniques, portant sur l'évaluation d'une chimiothérapie, la définition de la population incluse doit préciser s'il y a eu ou non une limite d'âge. Dans le cas d'une chimiothérapie cardio-toxique, il convient d'indiquer les critères cardiologiques d'exclusion de l'étude. Si la population étudiée concerne des malades, ce qui est le plus souvent le cas, il est indispensable que les critères sur lesquels on a fait le diagnostic de la maladie soient bien précisés. Toutes ces remarques semblent aller de soi, mais ne sont pas toujours évidentes en pratique. Ainsi, une étude prospective randomisée sur le traitement chirurgical des pancréatites aiguës biliaires avait été réalisée pour savoir s'il était préférable d'opérer dans les 48 heures ou de façon différée. Le protocole prévoyait trois critères d'inclusion :

– l'existence d'une douleur aiguë de type pancréatique dont le siège, les irradiations, les modalités d'apparition avaient été précisés ;
– une élévation de l'amylasémie au-dessus d'un certain seuil ;
– une lithiase biliaire reconnue sur un examen morphologique, en général une échographie abdominale.

Or, après inclusion d'une vingtaine de malades, les auteurs se sont aperçus que près de la moitié d'entre eux n'avaient pas de signes macroscopiques de pancréatite aiguë à l'intervention [1]. Les critères de diagnostic de pancréatite aiguë qui semblaient corrects et suffisants ne l'étaient pas pour deux raisons. D'une part, des lithiases biliaires sans pancréatite peuvent entraîner une élévation de l'amylasémie ; d'autre part, des coliques hépatiques peuvent donner des douleurs dont les caractères peuvent être similaires à ceux d'une pancréatite. Il eut été souhaitable, dans ce type d'étude, et pour ces raisons, d'exiger comme critère d'inclusion supplémentaire la preuve d'une pancréatite par la constatation d'une augmentation de volume du pancréas par un examen morphologique préopératoire, échographie ou scannographie par exemple.

Il convient encore d'indiquer si les sujets, qui répondent bien aux critères d'inclusion, ont été inclus dans l'étude de façon consécutive ou, dans le cas contraire, si des sujets qui auraient pu être inclus ne l'ont pas été en précisant leur nombre et les raisons de non-inclusion, même si elles peuvent paraître triviales. Ainsi, dans une étude prospective dont l'objectif était de comparer, chez des malades qui avaient un cancer de la tête du pancréas, les résultats de

l'écho-endoscopie et de l'écho-Doppler pour évaluer l'envahissement éventuel de la veine porte, des malades qui remplissaient bien tous les critères d'inclusion n'ont pas été inclus dans l'étude parce que l'appareil d'écho-Doppler était en panne. Ces exclusions doivent être indiquées avec leur(s) raison(s) car elles peuvent entraîner des biais dans l'analyse des résultats.

La période sur laquelle a porté l'étude doit encore être précisée. Les résultats peuvent différer si une étude a commencé en l'année 2005 ou bien a été réalisée à partir de l'année 2010. En effet, des changements parfois imperceptibles, de toute nature, ont pu survenir entre ces deux périodes.

Les critères d'inclusion et, le cas échant, d'exclusion ayant été précisés, il faut ensuite décrire la population qui a été retenue dans l'étude, par exemple l'âge moyen (avec l'intervalle de confiance ou les extrêmes), la répartition entre hommes et femmes, etc. dans la mesure où ces éléments descriptifs peuvent avoir un intérêt, c'est-à-dire sont pertinents.

Des animaux

Dans un travail portant sur des animaux, il convient de préciser l'espèce, la souche, l'âge, le sexe, le poids des animaux. Leur origine, leurs conditions d'élevage peuvent encore être utiles à connaître. Un laboratoire avait cru découvrir une souche de chats sujette à l'ostéoporose. Les rhumatologues avaient été très intéressés par ce « modèle animal » jusqu'au jour où ils se sont aperçus que les animaux qui leur étaient fournis avaient une ostéoporose due à une malnutrition sévère avant leur arrivée au laboratoire.

Des prélèvements

Tout travail portant sur des échantillons doit indiquer sur qui le prélèvement a été réalisé (être humain ou animal) et sur quoi (tissu, sang, sécrétion, etc.). Dans certains travaux, il est encore nécessaire de préciser la technique de prélèvement elle-même, ainsi que les conditions éventuelles de conservation si l'échantillon n'a pas fait l'objet d'un examen immédiat (congélation, fixation, milieu de culture, etc.).

Matériel d'étude clinique
Comment la population étudiée a-t-elle été sélectionnée ?

– *Critères d'inclusion*
En fonction des sujets eux-mêmes (âge, sexe, etc.).
En fonction de leur maladie.

– *Critères d'exclusion*
Nombre de sujets exclus.
Raisons de l'exclusion.

– *Divers*
Inclusions : consécutives, sinon pourquoi ?
Période sur laquelle a porté l'étude ?
Consentement éclairé des sujets ayant été inclus dans une étude prospective.

Description de la population : âge, sexe, etc.

Comment a-t-on travaillé ?

Une donnée fondamentale concerne la manière dont les données ont été recueillies :
- dans le temps, en différenciant les études rétrospectives, transversales, prospectives et longitudinales ;
- dans l'espace, en précisant s'il s'agit d'une étude unicentrique ou multicentrique.

Dans le temps

Examen rétrospectif de données

Il est possible de faire un travail sur l'examen rétrospectif de données recueillies avant la conception d'une étude, par exemple sur des dossiers plus ou moins anciens. L'inconvénient de ce type d'étude est que, par définition, le recueil des données n'a pas été établi dans la perspective de la réalisation d'un travail donné. De ce fait, certaines données peuvent manquer pour quelques sujets ou ne pas être pertinentes car la technologie a évolué pendant la période d'étude. Néanmoins, de telles études, si elles sont bien faites, sont utiles pour la connaissance de l'histoire naturelle de certaines maladies et sont presque indispensables avant d'élaborer des études prospectives.

Études transversales

Les études transversales (*cross-sectional* en anglais) consistent à recueillir des observations à une date donnée. Elles servent le plus souvent à quantifier l'importance d'un problème de santé dans une population donnée. Répétées dans le temps et dans les mêmes

conditions, elles permettent de suivre l'évolution, par exemple de survenue d'infections nosocomiales dans un établissement hospitalier.

Études prospectives

Les études prospectives recueillent les données au fur et à mesure de l'inclusion de nouveaux cas dans l'étude. Elles limitent les inconvénients des études rétrospectives. Leur principal inconvénient est souvent leur durée, qui dépend du rythme avec lequel il est possible de réaliser les inclusions.

Études longitudinales

Les études longitudinales consistent à suivre dans le temps des cohortes de sujets, par exemple en épidémiologie, de sujets exposés et non exposés à un facteur de risque potentiel.

Dans l'espace

Afin d'augmenter le nombre de cas inclus dans une étude rétrospective ou de réduire le temps des inclusions dans une étude prospective, il est possible de faire des études multicentriques, c'est-à-dire menées de façon concomitantes par plusieurs équipes différentes. Ces études ont comme autre avantage l'élaboration de protocoles, discutés en commun et qui doivent être particulièrement précis. En contrepartie, elles comportent un risque d'hétérogénéité qu'il convient de réduire le plus possible. Par exemple, dans un travail prospectif chirurgical multicentrique dont l'objectif est de comparer deux techniques chirurgicales, il est nécessaire de définir avec une grande précision les protocoles opératoires afin d'assurer une bonne homogénéité dans leur réalisation par les différents chirurgiens.

Un avantage majeur de ces études multicentriques, en pratique clinique, est que, étant réalisées par plusieurs participants, leurs conclusions sont plus largement extrapolables que les résultats d'une étude réalisée dans un seul centre, très spécialisé, avec un petit nombre d'opérateurs particulièrement entraînés.

Qui a fait quoi ?

Les études multicentriques posent, avec une particulière acuité, la question de leur publication. Il convient, en effet, qu'il y ait un (ou deux) maître(s) d'œuvre qui est (ou sont) « l'investigateur principal ». Celui-ci est, en général, à l'origine de l'idée du travail, de l'élaboration

du projet de protocole, du suivi des inclusions et de leur validation et, en fin d'étude, du recueil des données et de la rédaction du texte qui sera publié. Une partie de ces tâches peut être accomplie par des assistants de recherche clinique. Mais les différents participants seront d'autant plus actifs qu'ils n'ont pas de sentiment de frustration par rapport à cet investigateur. Pour ce faire, il y a tout intérêt à préciser, dès l'élaboration du protocole, les règles d'éventuelles publications et de communications à des congrès, du travail commun.

Il y a différentes façons de procéder.

L'une consiste à signer la publication sous le nom du groupe qui a réalisé l'étude multicentrique ; par exemple *Gastro-intestinal tumor study group* (GITS) ou *Veteran administration* (VA) ou Association française de recherche en chirurgie (FRENCH), etc. Une note en bas de première page ou en fin d'article précise alors le rôle de chacun : investigateur(s) principal(aux), participants par ordre alphabétique ou par importance décroissante de cas inclus dans l'étude, autres collaborateurs (statisticien, radiologue, anatomopathologiste, etc.).

Néanmoins, il est plus encourageant et plus motivant pour ceux qui ont travaillé le plus d'être mieux gratifiés : l'investigateur principal devient alors le premier signataire du travail, suivi des noms des quatre ou cinq principaux participants, puis du sigle du groupe ; le nom et les coordonnées des autres participants étant indiqués en note.

Signalons enfin que, dans le même état d'esprit, l'Association universitaire de recherche en chirurgie (l'AURC qui avait précédé FRENCH) avait prévu que les résultats d'un travail commun ne pourraient être présentés dans un congrès international sous forme de communication orale plus de deux fois par un même membre de l'association afin d'éviter que quiconque ne tire un profit personnel excessif d'un travail collectif.

Les statistiques

Les statistiques et les tests font l'objet des autres chapitres de cet ouvrage. Ils doivent toujours être précisés. En effet, des auteurs utilisent parfois des logiciels d'analyse et des tests statistiques inadaptés aux données qu'ils ont recueillies et qu'ils cherchent à interpréter.

Comment a-t-on travaillé ?

Type d'étude

– Recueil des données : rétrospectif, transversal, prospectif, longitudinal.
– Unicentrique, multicentrique.
– Recherche expérimentale (essai randomisé) ou observationnelle (transversale, cohorte, cas-témoin).

Statistique
– Dans des comparaisons, seuil des risques acceptés et nombre de sujets à inclure.
– Traitement de l'information.
– Tests d'inférence statistique, etc.

Ce que l'on a cherché à évaluer

Les travaux cliniques portent sur trois principaux types d'évaluation.

Les évaluations d'un « outil » diagnostique

Les évaluations d'un « outil » diagnostique doivent préciser toutes les données concernant cet « outil », qu'il s'agisse d'un symptôme, d'un signe clinique, d'un examen radiologique, isotopique ou biologique. Ainsi, dans une étude sur le diagnostic du pemphigus médicamenteux par un immunomarquage, il convient de préciser la technique de l'immunomarquage et de fabrication des réactifs qui ont été utilisés. S'il s'agit d'un examen radiologique, le type d'appareil et la référence du fabricant doivent être indiqués, les appareils et leurs performances évoluant avec les progrès technologiques.
La définition de la normalité n'est pas toujours évidente et mérite toujours d'être précisée. Par exemple, deux études sur la valeur des signes biologiques « anormaux » dans le diagnostic de métastases hépatiques ont montré des résultats assez différents. En fait, ces différences s'expliquaient par la raison suivante : dans une étude, les valeurs considérées comme anormales étaient celles indiquées par le laboratoire [2], alors que dans l'autre étude, le résultat était considéré comme anormal s'il était supérieur à la valeur moyenne plus deux écarts types chez les sujets inclus dans l'étude et qui n'avaient pas de métastases hépatiques [3].

Les évaluations thérapeutiques

Les évaluations thérapeutiques sont assez faciles à préciser pour un médicament : posologie, mode d'administration, horaires de prise. Le maximum de difficultés se rencontre dans la description et la réalisation des actes techniques dans une étude multicentrique pour assurer une homogénéité dans la réalisation de ce que l'on cherche à évaluer. Par exemple, dans une étude multicentrique néerlandaise sur le curage ganglionnaire dans le cancer de l'estomac, outre le protocole écrit, les chirurgiens disposaient d'un film sur la technique qu'ils devaient

appliquer [4]. De plus, les premiers malades que chaque chirurgien opérait n'étaient pas inclus dans l'étude pour éviter une hétérogénéité liée à ce que l'on appelle « la courbe d'apprentissage ». Dans ce type d'étude, il n'en reste pas moins que certains malades sont opérés par des chirurgiens qui sont de meilleurs opérateurs que d'autres. Les conséquences de ces facteurs d'hétérogénéité, appelés « effet centre » ou « effet opérateur », doivent être contrôlées en fin d'étude.

Les facteurs de pronostic et les facteurs de risque

Un troisième type d'études concerne les facteurs de pronostic et les facteurs de risque. Il peut s'agir d'études cliniques cherchant à évaluer des covariables qui sont liées à un bon ou à un mauvais pronostic. Il peut encore s'agir d'études épidémiologiques concernant des facteurs de risque d'apparition d'une maladie. Dans une étude épidémiologique sur les causes d'obésité, par exemple, il est nécessaire de préciser les variables qui sont étudiées et la manière dont elles sont mesurées : alimentaires, comportementales, génétiques, etc.

Ce que l'on cherche à évaluer

Un « outil » diagnostique qui peut être un symptôme, un signe clinique, un examen biologique, un examen radiologique, etc.

Un traitement qui peut être médical, chirurgical, par des agents physiques, etc.

Un facteur de pronostic d'une maladie ou **un facteur de risque** d'apparition d'une maladie, etc.

Critères de jugement

Assez curieusement, cette dernière partie du chapitre « Matériel et méthodes » d'un travail et plus encore d'un projet de recherche est parfois lacunaire, alors qu'elle devrait être aussi précise que les autre parties de ce chapitre [5].

Les critères de jugement diffèrent selon ce que l'on a cherché à évaluer. Si l'évaluation a porté sur la valeur d'un « outil » diagnostique, le critère de jugement est la présence ou l'absence de maladie dans la population sur laquelle cette évaluation a porté. En fait, ce « référentiel externe », cet examen de certitude, ce *gold standard*, n'est pas toujours évident. Si l'on cherche à estimer la valeur du Pet-Scan dans le diagnostic de métastases hépatiques, il faut savoir comment le diagnostic de métastases a été fait et surtout comment on a pu déterminer qu'il n'y avait pas de métastases ? En effet, une tumeur bénigne peut

simuler une métastase et des métastases de moins de quelques millimètres peuvent échapper à tout autre examen morphologique que celui qui est testé. En revanche, si l'on précise que les métastases ont été reconnues par un examen anatomopathologique et que l'absence de métastases a été confirmée avec un recul minimal de six mois, les choses deviennent plus convaincantes.

Dans un essai thérapeutique, le critère de jugement sera la survie ou le décès, la guérison ou la poursuite de la maladie, la récidive ou non, la durée d'hospitalisation, etc. Mais s'il s'agit d'une mortalité postopératoire, par exemple, s'agit-il de la mortalité au cours de l'hospitalisation qui suit l'intervention chirurgicale, même si elle survient deux mois et demi après, parce que le malade a fait des complications ou bien de la mortalité dans le mois qui a suivi cette intervention, même si l'opéré est sorti de l'hôpital ? Rien n'est simple. Tout doit être précisé. Bien souvent, les critères de jugement sont multiples. Dans une chimiothérapie pour tumeur solide, ce sera la survie, la régression tumorale (comment l'a-t-on mesurée ?), les contreparties hématologiques, digestives, la qualité de vie. Dans cet ensemble, il convient de distinguer le critère de jugement principal qui permettra finalement de conclure à l'efficacité ou non de l'intervention thérapeutique, des autres critères de jugement. C'est encore à partir du critère de jugement principal que l'on estime l'effectif des sujets qu'il est nécessaire d'inclure dans l'étude pour limiter le risque de deuxième espèce (cf. p. 92). Au bout du compte, la décision sera parfois difficile à prendre. Si une chimiothérapie « nouvelle » par rapport à une chimiothérapie « de référence » fait gagner une durée médiane de survie de cinq semaines au prix d'une mauvaise tolérance, exprimée par un pourcentage plus élevé de vomissements ou de leucopénies et d'une moins bonne qualité de vie, appréciée sur des critères précis, sera-t-il judicieux de proposer ou non un tel traitement au malade ? Évidemment, un abord purement théorique du problème serait d'expliquer les avantages et les inconvénients au malade. Mais est-il psychologiquement souhaitable de lui faire part de la gravité du pronostic d'autant plus, comme nous le verrons, que ce pronostic ne fait qu'exprimer des probabilités ? Enfin, si des études montrent qu'après tel type d'infarctus du myocarde, les probabilités de survie à cinq ans étaient de 60 %, il est impossible de prédire chez un malade déterminé s'il sera dans les 60 % de survivants ou dans les 40 % de patients qui vont décéder.

Dans l'étude d'un facteur de risque, le critère de jugement sera le risque qu'il convient de définir clairement ainsi que les données sur lesquelles l'apparition ou l'absence de survenue du risque ont été établies. Si ce facteur de risque concerne le pronostic d'une maladie, le critère de jugement sera la guérison, la survie ou bien, au contraire, le décès, la récidive.

Les critères de jugement

Dans un « outil » diagnostique : le référentiel externe.

Pour un traitement : la guérison, la survie, la récidive, etc. sans oublier les contreparties du traitement.

Il convient de bien distinguer :
– le critère de jugement principal ;
– les critères de jugement secondaires.

Pour un facteur de risque : dans un pronostic : la survie, la récidive, etc. En épidémiologie, la survenue ou non d'une maladie.

Références

1. Mackie CR, Wood RAB, Preece PE, Cushieri A (1995) A surgical pathology at early elective operation for suspected acute gallstone pancreatitis: preliminary report of a prospective clinical trial. Br J Surg 72: 179-81
2. Adloff M, Arnaud JP (1985) Étude prospective critique des différentes méthodes de détection des métastases hépatiques. Ann Gastroenterol Hepatol 21: 31-4
3. Molkhou JM, Lacaine F, Houry S, Huguier M (1989) Dépistage des métastases hépatiques des cancers digestifs. Place des dosages enzymatiques et de l'échographie. Presse Med 18: 1370-4
4. Bonenkamp JJ, Hermans J, Sasako M, *et al.* (1999) Extended lymph-node dissection for gastric cancer. N Engl J Med 340: 908-14
5. Chan AW, Altman DG (2005) Identifying outcomes reporting bias in randomised trials on PubMed : review of publications and survey of authors. BMJ 330: 753

Les variables qualitatives

2

Il existe plusieurs sortes de variables qualitatives (tableau I).

Tableau I – Les variables qualitatives.

Exemples
À deux modalités (ou dichotomiques)
– Vomissement : Oui/Non.
– Infection urinaire : Oui/Non.
À plusieurs modalités
– *Ordonnées.* Indice de masse corporelle*
• entre 18,5 et 24,9 (normal) :
• de 25 à 29 (surpoids) ;
• de 30 à 35,9 (obésité) ;
• ≥ 40 (obésité sévère).
– *Non ordonnées.* Infarctus du myocarde
• antérieur :
• antéro-septal ;
• postérieur.
* L'indice de masse corporelle, chez l'adulte, est égal au poids en kilogrammes divisé par le carré de la taille en mètres (il existe des corrections chez l'enfant en fonction de l'âge).

Les variables qualitatives sont dites « ordonnées » s'il existe un ordre naturel des modalités. Par exemple, dans un cancer de l'estomac, l'existence d'un envahissement ou non de la muqueuse, de la musculeuse, de la séreuse ou au-delà de la séreuse. Un autre exemple est le score METAVIR dans les hépatites chroniques [1].

Les variables qualitatives ne sont pas ordonnées s'il n'y a pas de relation d'ordre entre elles comme dans l'exemple du tableau I sur l'infarctus du myocarde.

Mesure

Dans un échantillon, par exemple chez un groupe de patients, une variable qualitative se mesure, comme chacun sait, par un pourcentage : sur 76 enfants, si huit ont eu la rougeole, le pourcentage

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

d'enfants ayant eu la rougeole est de huit sur 76, soit 10,5 % que l'on peut encore écrire 0,105.

En fait, ce pourcentage a été estimé sur cet échantillon de 76 enfants. Il est probable que, sur un autre groupe d'enfants, on aurait observé un pourcentage différent. Si l'on faisait des mesures sur un grand nombre d'échantillons, les valeurs des pourcentages observés se répartiraient selon une loi normale (cf. p. 17). Ces variations d'estimation de pourcentages d'un échantillon à l'autre suggèrent de rapporter aussi l'intervalle de confiance (*confidence interval* en anglais) qui donne une fourchette dans laquelle, à partir d'une mesure sur un échantillon, on estime que se situe la réalité. En général, cette fourchette est estimée de telle sorte qu'il y ait 95 chances sur 100 pour que la réalité se situe dans ses limites. C'est la « couverture » de l'intervalle de confiance (tableau II). On peut dire aussi que cet intervalle est « au risque de 5 % » ; c'est-à-dire qu'il y a 5 % de risque que la réalité se situe en dehors des limites de la fourchette qui a été estimée.

Tableau II – L'intervalle de confiance (approximation de la loi normale).

Exemple : sur un échantillon de 76 enfants, huit ont eu la rougeole (10,5 %).

p représente la proportion observée dans l'échantillon (dans notre exemple 10,5 % ou 0,105).

q est le complément $1 - p$ (soit 89,5 % ou 0,895).

n est l'effectif de l'échantillon (ici 76).

z_α est un coefficient dont la valeur dépend de l'intervalle de confiance que l'on souhaite calculer. Pour un intervalle de confiance à 95 %, la valeur de z_α est de 1,96.

L'intervalle de confiance à $(1 - \alpha)\%$ (IC ; confidence interval ou CI en anglais) se calcule ainsi :

$$\text{IC} = p \pm z_\alpha \sqrt{\frac{p \times q}{n}}$$

soit dans notre exemple :

$$\text{IC} = 0{,}105 \pm 1{,}96 \sqrt{\frac{0{,}105 \times 0{,}895}{76}} = 0{,}105 \pm 0{,}069$$

La valeur 0,069 représente la **précision de l'intervalle de confiance**.

À partir du pourcentage observé dans l'échantillon observé, l'intervalle de confiance à 95 % se situe donc entre (0,105 – 0,069), soit 0,036 et (0,105 + 0,069), soit 0,174.

Les valeurs de l'intervalle de confiance s'expriment alors ainsi : IC = [0,036 ; 0,174].

Quelques remarques

Pour être valide, cette approximation par la loi normale nécessite que np et nq soient ≥ 5.

On devine intuitivement que, plus l'échantillon est important (dans la formule du tableau II, le *n* du dénominateur), plus l'intervalle de confiance est petit, et réciproquement.
Si l'on souhaitait avoir une estimation de l'intervalle de confiance avec une probabilité supérieure à 95 %, par exemple 99 %, le coefficient z_α serait différent ; en l'occurrence, plus élevé (2,575) ; inversement, si l'on souhaitait se donner, par exemple, seulement 90 % de chances que l'intervalle de confiance contienne le vrai pourcentage, il serait moins élevé : 1,645 (tableau III).

Tableau III – Valeurs du coefficient z_α en fonction de la probabilité que l'on souhaite se donner pour le calcul de l'intervalle de confiance.

Probabilité souhaitée	90 %	**95 %**	99 %	99,5 %	99,9 %
Valeurs du coefficient z_α	1,645	**1,960**	2,576	2,807	3,291

Dans tout travail faisant état de variables qualitatives, mesurées par un pourcentage de leur présence dans l'échantillon étudié, ce dernier doit être assorti de son intervalle de confiance à 95 %. On s'aperçoit alors, bien souvent, qu'une ou des décimales dans l'expression d'un pourcentage sont dérisoires. Ces décimales suggèrent que les auteurs n'ont guère réfléchi à la notion d'intervalle de confiance. Dans notre exemple de rougeole, la décimale du pourcentage observé, 10,5 %, est dérisoire par rapport à l'étendue de l'intervalle de confiance qui va de 4 % (très précisément 3,6 %) à 17 % (très précisément 17,4 %), ce qui suggère que l'on aurait pu arrondir à 11 %.
On se rend encore compte que, pour un même nombre de cas étudiés (*n*), plus les pourcentages se rapprochent de 50 %, plus *pq* est grand et plus l'intervalle de confiance est important. Dans les sondages d'opinion qui, pour des raisons de coût, portent sur environ un millier de personnes, si le nombre d'opinions en faveur de A est de l'ordre de 45 % (et en faveur de B de 55 %), la précision est de l'ordre de 3 % et IC = [42 ; 48].

Référence

1. The French METAVIR cooperative study group (1994) Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. Hepatology 20: 15-20

Les variables quantitatives 3

Comme leur nom l'indique, les variables quantitatives servent à représenter des quantités. Ces variables, lorsqu'elles prennent des valeurs réelles, c'est-à-dire correspondent à un *continuum* de valeurs exprimées avec des décimales, comme le poids en kilogrammes (63,38 kg), la taille en mètres (1,76 m) sont appelées des variables continues[1]. Entre deux valeurs proches l'une de l'autre, il peut toujours en exister une troisième.
Si des variables quantitatives ont des valeurs entières comme l'âge en années, un nombre de journées d'hospitalisation ou, chez un malade qui a de la diarrhée, le nombre de selles par 24 heures, ces variables sont appelées discrètes.

Les variables quantitatives. Exemples

Variables continues : créatinémie en micromoles par litre exprimée avec décimales.

Variables discrètes : nombre de grossesses.

Variables continues. Distributions. Représentations graphiques

Distribution

La distribution des variables continues suit, souvent, une courbe en cloche. La loi normale, au sens biostatistique du terme, encore appelée loi de Laplace-Gauss[2] peut alors être utilisée. Il y a d'autres lois pour

1 Rappelons à ce propos 1) que les abréviations internationales des unités de mesure ne peuvent être utilisées qu'après un nombre ; 2) qu'elles doivent être écrites en lettres minuscules avec quelques exceptions, (IC pour intervalle de confiance, DS pour déviation standard, L pour litre, Gy pour Gray, Pa pour Pascal) ; 3) qu'elles sont invariables ; 4) et que l'abréviation de minutes est min et non mn comme on le voit écrit trop souvent.

2 Pierre-Simon, marquis de Laplace (1749-1817) était un mathématicien, physicien et astronome français. Carl Friedrich Gauss (1777-1865) exerçait dans les mêmes disciplines et était allemand.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

variables continues, comme la loi log-normale qui est souvent appropriée pour les dosages biologiques, ou la loi de Weibull pour la survie. La distribution des variables quantitatives discrètes peut se rapprocher d'une distribution normale, mais, par définition, elles suivent des lois pour variables discrètes comme la loi binomiale ou la loi de Poisson[3].

Représentation et mesure

Si l'on veut représenter les valeurs d'une variable quantitative continue en portant en abscisse des valeurs de la variable (par exemple le taux sanguin d'acide urique sérique pour 100 mL chez des sujets sains) et en ordonnées le nombre de sujets qui ont une valeur donnée de ce taux d'acide urique, on obtiendrait ce que montre la figure 1. En effet, la probabilité d'avoir exactement la même valeur chez deux sujets est faible. Elle serait même nulle si l'on était capable de mesurer les valeurs avec une précision aussi grande que possible. C'est, du reste, ce qui définit une variable quantitative continue. En pratique, on calcule plutôt un tableau de fréquence (ou de distribution) en comptant les valeurs dans des classes de même largeur comme le montre le tableau I.

Tableau I – Distribution de la concentration d'acide urique sérique chez 267 hommes sains [1].

Taux d'acide urique (mg/100 mL)	Nombre de sujets	Fréquence relative	Fréquences relatives cumulées (%)
3,0 – 3,4	2	0,8	0,8
3,5 – 3,9	15	5,6	6,4
4,0 – 4,4	33	12,3	18,7
4,5 – 4,9	40	15,0	33,7
5,0 – 5,4	54	20,2	53,9
5,5 – 5,9	47	17,6	71,5
6,0 – 6,4	38	14,2	85,8
6,5 – 6,9	16	6,0	91,8
7,0 – 7,4	15	5,6	97,4
7,5 – 7,9	3	1,1	98,5
8,1 – 8,4	1	0,4	98,9
8,5 – 8,9	3	1,1	100,0

3 Siméon Poisson (1781-1840) était un mathématicien français.

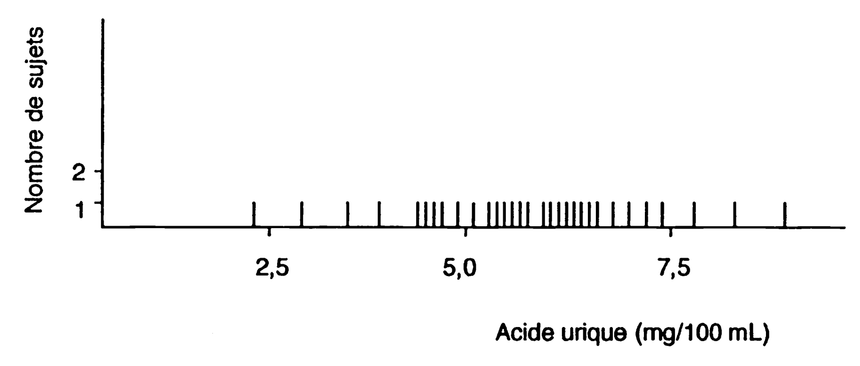


Fig. 1 – Représentation graphique d'une variable quantitative continue dite « *rug-plot* ».

Les valeurs observées peuvent alors être représentées sous forme d'histogramme (fig. 2).

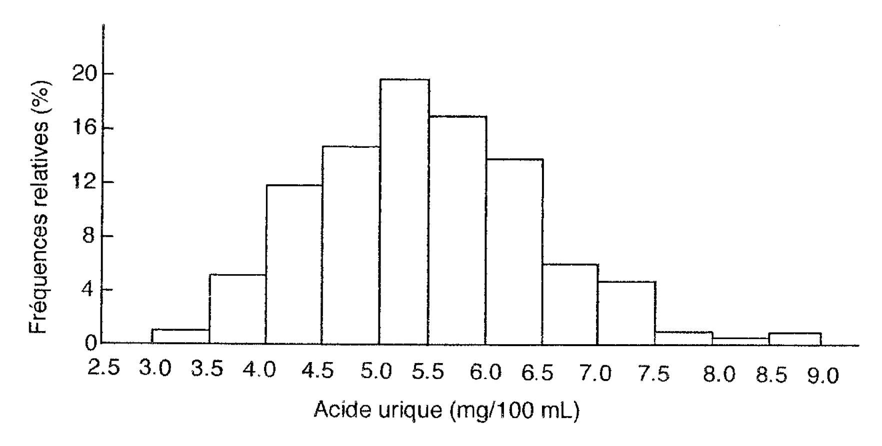


Fig. 2 – Histogramme représentant la distribution des concentrations de l'acide urique sérique dans une population de 267 hommes sains (d'après Morton *et al.* [1]).

Certains histogrammes sont parfois représentés avec des largeurs de colonnes qui diffèrent d'une colonne à l'autre (fig. 3). Dans ce cas, c'est la surface de la colonne qui doit correspondre à l'effectif du sous-groupe représenté, et non à sa hauteur. Dans la figure 3, la première colonne correspond à deux malades qui ont entre 20 ans et 40 ans, la seconde à trois malades qui ont entre 40 ans et 50 ans, la troisième colonne à quatre malades et la dernière colonne à quatre malades.

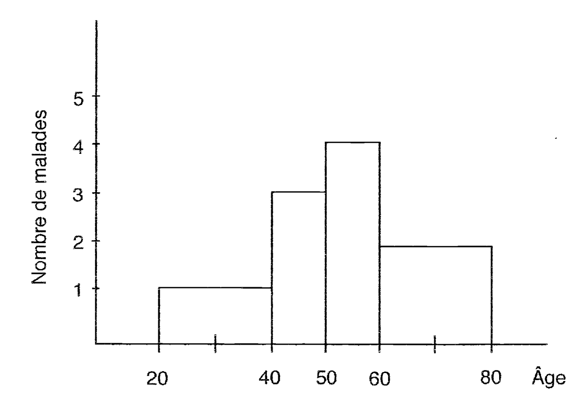


Fig. 3 – Histogramme avec des largeurs de colonnes différentes : les effectifs correspondent à la surface de la colonne et non à sa hauteur. Il y a deux sujets correspondant à la première colonne, trois dans le deuxième, quatre dans la troisième, et quatre dans la quatrième.

Il est encore possible de faire une courbe de fréquences relatives cumulées dont les valeurs sont indiquées dans la dernière colonne du tableau I (fig. 4). C'est ce que l'on appelle la fonction de répartition. Dans cette courbe, l'axe des abscisses représente les valeurs de la variable étudiée, et l'axe des ordonnées correspond au nombre de cas cumulés. La valeur correspondant à 50 % des fréquences relatives cumulées (encore appelée le 50ᵉ percentile) est la valeur médiane dans cette population.

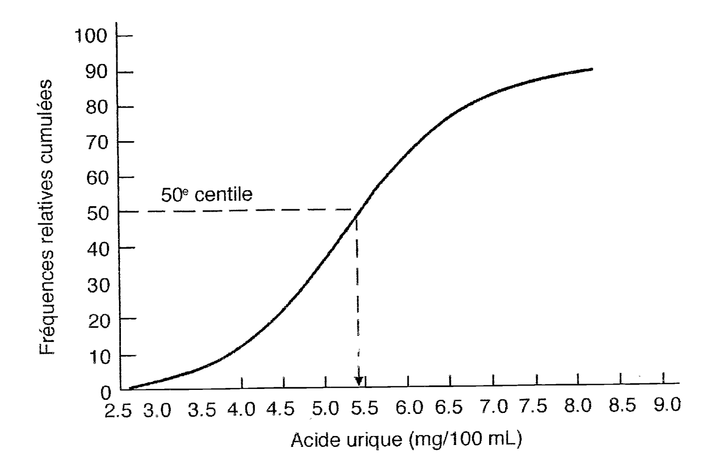


Fig. 4 – Fréquences relatives cumulées (encore appelées fonction de répartition) des concentrations de l'acide urique sérique dans une population de 267 hommes sains (d'après Morton *et al.* [1]).

Dans la figure 2 des histogrammes en fonction du taux d'acide urique, si l'on avait fait des mesures sur plusieurs milliers de sujets et que l'on avait pris des largeurs de colonnes correspondant non pas à des écarts de 0,5 mg/100 mL d'acide urique, mais de 0,1 mg/100 mL, on aurait eu une figure qui se rapprocherait d'une courbe de Laplace-Gauss (fig. 5).

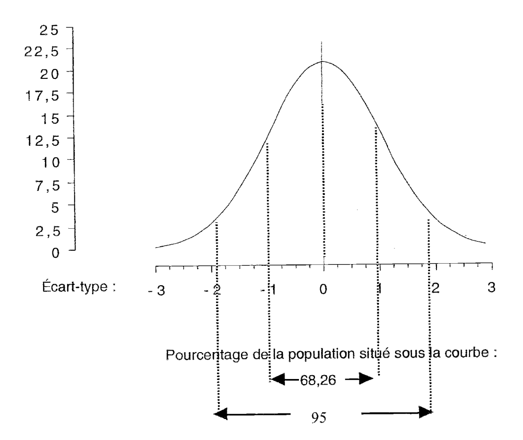


Fig. 5 – Courbe de Laplace-Gauss. Distribution normale et pourcentages de la population sous la courbe (aires sous la courbe).

Mesures descriptives. La loi normale (Laplace-Gauss)

Il existe plusieurs outils de mesures descriptives des variables quantitatives.

La moyenne (*mean* en anglais)

La moyenne arithmétique (désignée par la lettre m) de plusieurs variables quantitatives est égale à la somme des valeurs observées divisée par le nombre de mesures qui ont été faites (tableau II).

Tableau II – La moyenne arithmétique m.

$$m = \sum x_i / n$$

où :
– m est la moyenne (dans notre exemple 7 mois) ;
– x_i sont les valeurs observées chez chaque sujet (dans notre exemple 8, 10, 5, etc.) ;
– n est le nombre de sujets étudiés (dans notre exemple 10).

Si, sur dix malades atteints d'un cancer de très mauvais pronostic, les durées de survie en mois sont de 8, 10, 5, 12, 5, 4, 7, 6, 8 et 5 mois, la durée moyenne de survie de ce groupe de malade est de :

$$\frac{8+10+5+12+5+4+7+6+8+5}{10} = 7 \text{ mois.}$$

Si l'effectif de l'échantillon est faible, comme dans cet exemple, il suffit d'un événement inusuel pour changer notablement la moyenne. Par exemple, si la durée de vie du dernier malade avait été de 35 mois au lieu de 5 mois, le calcul montre que la durée moyenne de survie serait de 10 mois au lieu de 7 mois. Il convient ainsi de se méfier d'un résultat exprimé par une moyenne lorsque le nombre de mesures qui a permis son calcul est petit.

Un autre inconvénient de la moyenne est qu'il faut attendre que tout ce que l'on cherche à mesurer puisse l'être, c'est-à-dire la survenue du dernier élément : dans notre exemple, il faut attendre que tous les malades soient décédés pour pouvoir calculer la durée moyenne de survie de cette population.

La médiane

La médiane est la valeur pour laquelle 50 % des mesures sont plus grandes et 50 % plus petites. Dans l'exemple qui a été pris, la médiane est la valeur observée entre le 5e et le 6e malade, c'est-à-dire entre 6 mois et 7 mois (par convention, c'est la moyenne de ces deux valeurs, soit 6,5 mois lorsque l'on a un nombre pair d'observations). Lorsque la distribution des valeurs est normale (courbe de Laplace-Gauss), la médiane et la moyenne se confondent (fig. 5).

La médiane a l'avantage sur la moyenne de pouvoir être estimée sans attendre que tous les événements se soient produits : dans notre exemple, il suffirait que la moitié des malades soient décédés.

La variance et l'écart-type (*standard deviation* en anglais, *SD*)

La variance mesure la dispersion de la distribution des valeurs autour de la moyenne (et ceci, aussi bien dans la loi de Laplace-Gauss que dans la loi binomiale ou que dans la loi de Poisson) (tableau III). Elle est désignée par s^2.

L'écart-type est la racine carrée de la variance. Il est encore appelé déviation standard et désignée par la lettre s. De même que pour la variance, un écart-type faible signifie que les valeurs observées sont peu dispersées autour de la moyenne. Inversement, un grand écart-type traduit une dispersion importante (fig. 6).

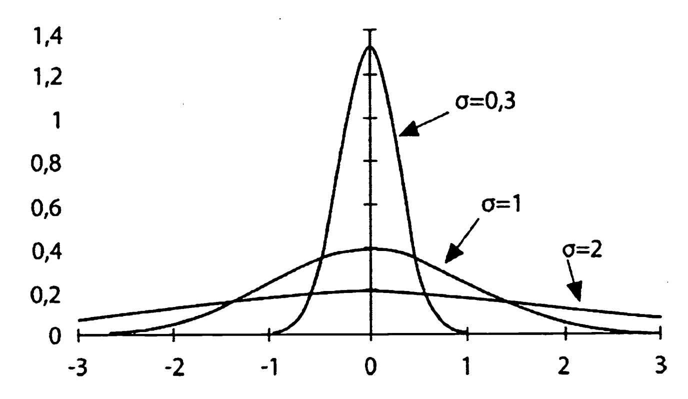


Fig. 6 – Courbes de Laplace-Gauss pour des valeurs différentes d'écarts-types (σ).

Tableau III – Variance et écart-type.

Pour chaque valeur observée (x), il est possible de calculer la différence ou écart (d) avec la moyenne (m) : $d = x - m$. L'estimation de la variance (s^2) est la moyenne des carrés des écarts autour de la moyenne.
Variance $s^2 = \frac{\sum (x-m)^2}{n-1}$ ou $\frac{\sum d^2}{n-1}$ ou encore $\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$
L'écart-type est la racine carrée de la variance : $s = \sqrt{s^2}$.

La moyenne plus ou moins *s* englobe 68,26 % de la population (fig. 5). La moyenne plus ou moins 1,96 × *s*, 95 % de la population[4]. Les deux paramètres que sont la moyenne et l'écart-type (ou la variance) suffisent à caractériser la loi de probabilité de distribution des valeurs qui suivent une loi normale.
La valeur de l'écart-type *s* ne dépend pas de la taille de l'échantillon contrairement à l'erreur standard de la moyenne *(standard error of the mean SEM)*. Cette dernière est égale à $s/\sqrt{n}$.

Moyenne et médiane

Comme il est dit plus haut, lorsqu'une distribution est symétrique, sa moyenne et sa médiane sont confondues. Dans le cas contraire, on peut trouver des exemples comme dans la figure 7, où à médiane semblable correspondent des moyennes différentes et vice-versa.

4 Si l'on arrondi 1,96 à 2, l'intervalle englobe 95,44 % de la population.

Moyenne et médiane sont deux outils de mesure qui permettent de communiquer de manière simple quelle est la valeur « typique » d'une observation. La moyenne est un bon outil descriptif lorsque la distribution des valeurs est symétrique. Dans les autres cas, on pourra préférer la médiane. Cette appréciation est cependant avant tout visuelle. Mais souvent les deux quantités donneront raisonnablement la même idée des données, comme une médiane de survie de 4 mois ou une durée moyenne de survie de 5 mois.

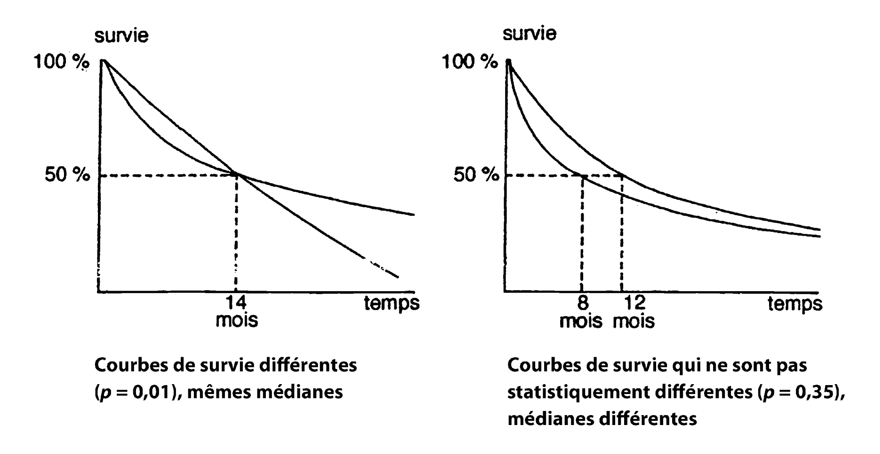


Fig. 7 – Exemples montrant que des courbes de survie différentes peuvent avoir la même médiane de survie et que des survies assez similaires peuvent avoir des médianes différentes.

La loi de Laplace-Gauss permet d'estimer des probabilités comme l'indique le tableau ci-dessous.

Loi de Laplace-Gauss

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$p(x)$ est la probabilité d'observer une valeur dans un intervalle infinitésimal autour de x.

σ est l'écart-type.

μ est la moyenne arithmétique.

Rappelons que e indique une exponentielle.

Dans une distribution normale, les valeurs sont symétriques par rapport à la moyenne (μ). On dit encore que la distribution est réduite si sa variance est égale à 1 et qu'elle est centrée si sa moyenne est égale à 0. Dans une distribution centrée, réduite, les probabilités sont donc définies par la formule simplifiée :

Loi de Laplace-Gauss (distribution centrée réduite)

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

En pratique, on peut estimer la probabilité d'événements au-dessus ou au-dessous d'une valeur, en calculant l'écart réduit z, en divisant son écart par rapport à la moyenne, par l'écart type σ, puis en rapportant la valeur calculée z_0 à une table qui donne cette probabilité (tableau IV).

Loi standardisée

$$z = \frac{(x - \mu)}{\sigma}$$

σ indique « la vraie valeur » de l'écart-type.

Tableau IV – Table de la loi normale réduite.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Par exemple, si la distribution d'une valeur biologique normale (poids d'un nouveau-né) a une moyenne de 3,3 kg et un écart-type de 0,5, la probabilité que des valeurs se situent au-dessous de 2,5 kg donne z égal à 2,5 – 3,3/0,5, soit 1,6. On cherche cette valeur (au signe près) dans la première colonne et ligne du tableau IV, en décomposant en décimale (ici 1,6) et centésimale (ici 0,00). À l'intersection se trouve la valeur 0,9452 qui donne la probabilité qu'une loi normale centrée soit inférieure à 1,6. En exploitant la symétrie de la loi normale, on en déduit qu'il y avait 1 – 0,9452 = 0,055 soit 5,5 % de chances d'observer une valeur plus petite que – 1,6 et donc un poids inférieur à 2,5 kg ; autrement dit, que cet enfant est au 5e percentile des poids de naissance. Si l'on voulait estimer la probabilité que des valeurs se situent entre 2,5 et 3,0 kg, ce qui peut s'exprimer par $p\ (2{,}5 \leq x \leq 3)$. De même que nous avons fait le calcul pour un poids de 2,5 kg, le calcul pour un poids de 3 kg donne (3,0 – 3,3)/0,5, soit 0,6. Comme précédemment, le tableau IV donne la probabilité d'avoir une valeur inférieure à 3 de 1 – 0,725 La probabilité que le poids se situe entre 2,5 kg et 3 kg est donc de 0,945 – 0,725, soit 0,220 ou 22 %.

Des calculs similaires pourraient être faits pour toutes les variables quantitatives continues, des examens biologiques par exemple, qui suivent une loi normale.

La loi binomiale

Lorsque la variable d'intérêt compte un nombre de « succès » parmi n tentatives semblables, il s'agit d'une variable binomiale. C'est, par exemple, dans des familles de deux enfants dont les parents ont une anomalie génétique autosomique récessive, le cas de distribution des tares observées chez les enfants, certains enfants ayant la tare, d'autres non. Dans certaines familles, les deux enfants peuvent ne pas avoir hérité de la tare. Dans d'autres familles, ce sera un enfant, dans d'autres encore les deux enfants. L'expression graphique d'une loi binomiale est celle d'un diagramme en bâtons (fig. 8). Il est possible de calculer la moyenne et l'écart-type d'une distribution binomiale. En revanche, ces deux paramètres, contrairement à ce que nous avons indiqué pour la loi normale, ne suffisent pas à décrire une distribution binomiale.

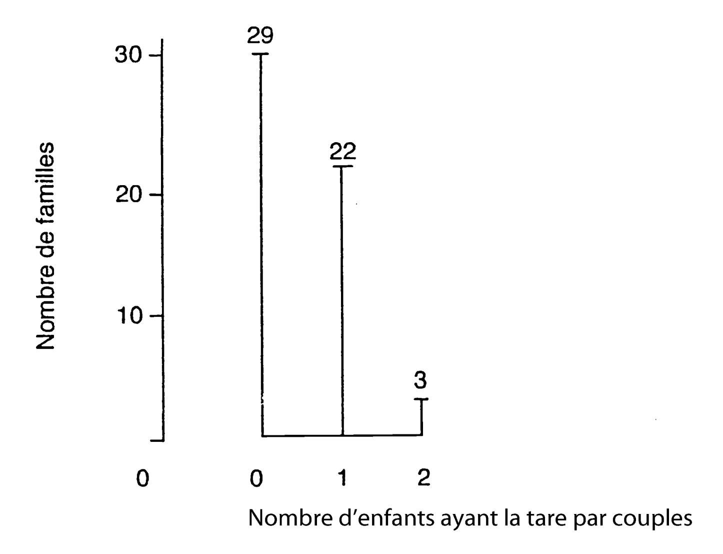


Fig. 8 – Expression graphique d'une distribution de variables quantitatives discontinues qui suivent une loi binomiale.

Exemple

L'utilité de la loi binomiale est, comme toute loi de distribution, de permettre le calcul des probabilités. Elle sert aussi à élaborer des tests statistiques exacts pour des variables qualitatives comme le test de Fisher (cf. page 84). Cette utilisation fait appel à deux notions : celle de factorielle et celle de combinatoire.

La *factorielle* d'un nombre est le résultat de la multiplication de tous les nombres entiers égaux et inférieurs à ce nombre. Par exemple, la factorielle de 6 est égale à 6 × 5 × 4 × 3 × 2 × 1, soit 720. Ceci s'écrit 6 ! = 720, le « ! » signifiant « factorielle ». Par convention, 0 ! = 1.

Une *combinatoire* (*C*) est le nombre de façons d'avoir *k* événements, parmi *n*. Cela s'écrit C_n^k[5]. Par exemple, dans le Tournoi de rugby des six nations représentées par l'Écosse, le pays de Galles, l'Angleterre, l'Irlande, la France, et l'Italie, les équipes jouant deux à deux l'une contre l'autre (*k*), la combinatoire C_n^k est le nombre de matchs nécessaire pour que chacune des six équipes (*n* = 6) rencontre chacune des autres équipes (*k* = 2). Cela peut se calculer assez facilement dans notre exemple, mais devient d'autant plus difficile que le nombre d'événements est important. La formule est donnée dans le tableau V.

5 On note plutôt actuellement une combinatoire en la faisant figurer entre parenthèses : $\binom{n}{k}$.

Tableau V – Formule générale d'une combinatoire et exemple avec le Tournoi de rugby des six nations.

$$C_n^k = \frac{n}{k!(n-k)} \text{, soit dans notre exemple :}$$

$$\frac{6!}{2!(6-2)} = \frac{6\times5\times4\times3\times2\times1}{2\times1(4\times3\times2\times1)} = \frac{720}{48} = 15$$

Dans le Tournoi des six nations, il est donc nécessaire d'organiser 15 matchs pour que chaque équipe rencontre l'autre.

Voyons maintenant un exemple d'application de la loi binomiale. Supposons qu'un examen soit doté de valeurs « normales » qui correspondent à 95 % de la population saine, en bonne santé. Ces limites n'incluant que 95 % (q) des sujets normaux, les 5 % (p) restants auront une valeur hors de ces limites. Si on fait plusieurs examens (n) indépendants entre eux à un sujet normal, la loi binomiale permet de calculer la probabilité que k de ces examens soient « anormaux », c'est-à-dire sortent de la fourchette des 95 % (tableau VI).

Tableau VI – Exemple d'application de la loi binomiale.

Les données

Les seuils adoptés dans un examen biologique n'incluent que 95 % des valeurs des sujets sains (p) et 5 % des examens ont un résultat qui sort des valeurs considérées comme normales ($p = 1 - q$).

Si l'on réalise cinq examens biologiques (n = 5) indépendants (non liés entre eux) chez un sujet normal, la probabilité que l'un de ces examens soit « anormal » ($k = 1$) peut être estimée par la loi binomiale.

L'application de la loi binomiale donne le résultat :

$p\,(X = k) = C_n^k \times p^k \times q^n$, soit : $C_5^1 \times p^1 \times q^5$, ce qui donne : $C_5^1 \times 0{,}05^1 \times 0{,}95^5 = 0{,}19$.

Cet exemple montre que si l'on demande cinq examens biologiques indépendants, dont les valeurs « normales » correspondent à 95 % des cas, il y a 19 % de chances que le résultat de l'un de ces cinq examens soit apparemment « anormal ». Il serait facile de démontrer que si l'on demandait, non plus cinq, mais dix examens indépendants entre eux, la probabilité que l'un d'entre eux soit apparemment « anormal » s'élèverait à 30 %. Cela prouve l'absurdité de ces bilans qui sont trop souvent demandés en pratique, notamment hospitalière, ou qui sont réalisés au nom de la facilité de leurs dosages. Cette notion biostatistique montre, *a contrario*, que des examens biologiques doivent être demandés en fonction d'hypothèses cliniques préalablement formulées afin d'en réduire le nombre et partant, le risque de ces « faux positifs ».

La loi binomiale étant moins connue et moins utilisée que la loi de Laplace-Gauss, nous en donnons un autre exemple d'application : supposons qu'un traitement anticancéreux de référence soit connu pour sa neurotoxicité : 10 % des malades traités développent une neuropathie périphérique. Supposons qu'un autre traitement, de même efficacité, ait été mis au point, et que l'on espère cet autre traitement moins neurotoxique que le traitement de référence. Si l'on traite 50 malades avec ce nouveau médicament, dans les mêmes conditions que le traitement de référence, et qu'il n'apparaît aucun cas de neurotoxicité, il convient d'interpréter ce résultat. C'est ce que permet la loi binomiale. En effet, le nombre de cas de neuropathie peut être décrit par une variable aléatoire X qui suit une loi binomiale, car on compte le nombre de malades qui n'ont pas eu de neuropathies (k), et ceux qui auraient eu une neuropathie ($n - k$), zéro dans notre exemple, chaque malade ayant la même probabilité (p) d'avoir une neuropathie. On peut, grâce à la loi binomiale, calculer la probabilité que l'on aurait eue, avec l'anticancéreux de référence, d'observer sur 50 malades (n), k cas de neuropathie. Cette probabilité s'écrit on l'a vu (tableau VI), $P(X = k) = C_n^k \times p^k \times q^n$ où k est le nombre de cas observés (zéro dans notre exemple).
Le calcul donne :

$$p(X = 0) = C_{50}^{0} \,.\, 0{,}1^0 \,.\, 0{,}9^{50}, \text{ soit } 0{,}005.$$

Autrement dit, avec le traitement de référence, il n'y avait que cinq chances sur 1 000 de n'observer aucune neuropathie sur les 50 malades traités. On pourra donc penser que le nouveau traitement ne donne pas autant de neuropathies que le traitement de référence. Comme on le verra, on a en fait bâti ici un test statistique non paramétrique exact, dont on reverra les principes ultérieurement (cf. page 84).
Lorsque la taille d'une population dans laquelle on mesure une variable quantitative discontinue est suffisante (en pratique dès que np et nq sont > 5), on peut faire l'approximation que la distribution se rapproche d'une loi normale et la probabilité en est grande, ce qui simplifie les calculs et permet d'utiliser des tests paramétriques (cf. page 84) dans les comparaisons.

La loi de Poisson

La loi de Poisson est, elle aussi, utilisée pour des variables discrètes. Son application la plus courante est l'approximation de la loi binomiale lorsque les événements sont rares et l'échantillon suffisamment grand.
La formule générale en est :

$$p(X = k) = \frac{\lambda^k}{k\,!} \cdot e^{-\lambda}$$

où λ est le paramètre de la loi de Poisson. Il est égal à la moyenne de *X* ainsi qu'à sa variance (la moyenne et la variance étant égales dans la loi de Poisson). Si la loi de Poisson est utilisée en approximation de la loi binomiale, on a $\lambda = n \, . \, p$; *n* étant l'effectif de la « population » étudiée et *p* la probabilité de survenue de l'événement, *e* est la fonction exponentielle.
Dans notre exemple concernant les examens biologiques (tableau VI), *n* était égal à 5, *p* était égal à 5 %. On a donc (tableau VII) :

Tableau VII – Exemple d'application de la loi de Poisson en reprenant les données du tableau VI.

$$p(X=k)=\frac{\lambda^k}{k!}\cdot e^{-\lambda}=\frac{(5\times 0,05)^1}{1!}\times e^{-0,25}=0,25\times 0,78=0,195$$

La probabilité que l'on a calculée avec la loi de Poisson est très proche de celle calculée avec la loi binomiale qui était de 0,19. On peut en effet démontrer que lorsque *p* est suffisamment petit devant *n*, la loi de Poisson se rapproche de la loi binomiale. La loi de Poisson a l'avantage d'être mathématiquement plus simple que la loi binomiale.
Autre illustration, dans notre exemple concernant le risque de neuropathie, on aurait :

$$p(X=0)=\frac{5^0}{0!}\cdot e^{-5}=1\times 7\cdot 10^{-3}=0,007.$$

Un autre exemple d'utilisation de la loi de Poisson est celui du comptage des colonies dans une boit de Pétri. Si une suspension bactérienne contient 5 000 bactéries par litre et que l'on ensemence une boîte de Pétri à partir de cette suspension, à raison de 1 cm^3 de solution, la probabilité qu'il n'y ait pas de colonie ensemencée est :

$$p(X=0)=\frac{5^0}{0!}\cdot e^{-5}$$

et la probabilité qu'il y ait au moins une colonie par boîte de Pétri est de :
$P(X > 0) = 1 - P(X = 0) = 1 - 0,0067 = 0,9933.$

Référence

1. Morton RF, Hebel JR (1990) Épidémiologie et biostatistique. Une introduction programmée. Doin, Paris, p 68-9

Les variables censurées

4

La censure intervient lorsque l'on n'observe pas exactement la valeur de la variable à laquelle on s'intéresse. L'exemple type est celui de la survie, ce qui requiert la survenue de l'événement « décès ». Par extension, cette notion de « décès » peut s'appliquer à tout événement non récurrent qui survient dans le temps, par exemple, l'apparition d'une récidive ou d'une métastase dans un cancer. La notion de « survie » s'applique alors au temps écoulé sans l'apparition d'une récidive ou d'une métastase. Les courbes qui en résultent décroissent avec le temps, comme une courbe de survie. Néanmoins, on a plus souvent tendance à faire figurer les taux de récidives ou d'apparition d'un événement pathologique, par une courbe croissante, ces taux n'étant que le complément des taux sans récidives. Ces courbes croissantes d'incidences cumulées sont également appropriées lorsque l'on s'intéresse à une cause de mortalité, mais qu'il peut y avoir d'autres causes (mortalités compétitives).
Les notions de temps jusqu'à l'événement nécessitent la définition d'un temps zéro (date d'origine) qui correspond à un même événement chez chaque personne, par exemple, la date de diagnostic ou de début d'un traitement.
De même qu'une variable quantitative peut être transformée en variable qualitative, on peut transformer une variable censurée en variable qualitative : pourcentage de patients survivants à cinq ans. Pour ce faire, il faut que le suivi minimum pour chaque patient soit au moins de cinq ans. Une variable censurée peut devenir simplement quantitative si la durée de survie est obtenue pour chaque individu dans une population. La transformation de variables censurées en variables quantitatives ou qualitatives fait perdre de l'information et peut introduire des biais.

Définitions

Des définitions sont importantes à préciser.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

La date d'origine. Il s'agit de la date de l'entrée dans l'étude d'un sujet comme la date de survenue d'un infarctus du myocarde ou d'une intervention chirurgicale.

La date des dernières nouvelles. Il peut s'agir ou bien de la date à laquelle le sujet a été vu la dernière fois ou bien de la date de la survenue de l'événement que l'on cherche à évaluer comme la date d'un décès ou de l'apparition d'une récidive.

Le temps de participation à l'étude. Il s'agit du délai entre la date des dernières nouvelles et la date d'origine.

La date de point. C'est la date à laquelle on fait le bilan des dernières nouvelles pour l'ensemble de la population que l'on étudie. L'idéal est que la date de point se confonde avec la date des dernières nouvelles pour les sujets en vie (ou sans récidive).

Les « perdus de vue ». Ce sont des malades en vie (ou sans récidive) lors de la date des dernières nouvelles si elle est inférieure à la date de point. Cette définition de l'antériorité peut dépendre de l'histoire naturelle de ce que l'on cherche à évaluer sur l'échantillon qui est étudié. Par exemple, lors d'une maladie grave dans laquelle la durée de survie est habituellement limitée à quelques mois, un malade qui n'a pas été revu dans le mois qui précède la date de point doit être considéré comme perdu de vue. Au contraire, lors d'une affection moins grave, dans laquelle le taux de survie à cinq ans est de l'ordre de 60 %, un malade pourra n'être considéré comme perdu de vue que si l'intervalle entre la date des dernières nouvelles et la date de point excède six mois.

Bien entendu, un des critères de qualité d'une étude est le faible pourcentage de perdus de vue. Sinon, des biais dans les résultats peuvent s'introduire, les sujets perdus de vue ne se comportant pas forcément comme les autres malades (tableau I). Compte tenu de la masse d'information dont on dispose, on peut recommander d'éviter de lire des articles dans lesquels le pourcentage de perdus de vue n'est pas précisé ou si ce pourcentage dépasse 10 %. Dans ce cas, il convient en effet d'interpréter les résultats avec la plus grande prudence.

Les exclus-vivants ou exposés au risque. Ce sont des sujets qui ne sont pas décédés au moment des dernières nouvelles (méthode de Kaplan-Meier) ou dans le dernier intervalle allant jusqu'à la date des dernières nouvelles (méthode actuarielle). Les exclus-vivants comportent les sujets perdus de vue et les sujets vivants à la date de point. Même si leur temps de participation à l'étude est relativement bref, ces sujets sont pris en compte car les résultats les concernant contribuent à l'acquisition de la connaissance. Ainsi, le fait de savoir qu'un patient a vécu six mois ou un an sans récidive ni métastase de son cancer apporte une information qui, bien que limitée, doit être prise en compte.

Tableau I – Les « perdus de vue ». Exemple de problèmes qu'ils posent.

Les données

Une étude avait porté sur les résultats d'une technique chirurgicale de traitement de hernies inguinales. Elle avait inclus 280 malades opérés depuis plus de deux ans. Il a été observé 14 récidives. Trente-quatre malades avaient été perdus de vue, six malades étaient décédés dans les deux ans, de cause sans rapport avec l'intervention.

Commentaires

Les auteurs estimaient leur taux de récidives à 14/280, soit 5 % (il aurait été souhaitable qu'ils indiquent aussi l'intervalle de confiance pour éviter d'avoir à le calculer). En fait :

– Si le calcul du taux de récidive avait été fait chez les 240 opérés qui avaient deux ans de recul, soit 14/240, on aurait trouvé 5,8 % de récidives.

– Mais, si tous les opérés perdus de vue ou décédés avaient fait une récidive, il y aurait eu 14 + 40 récidives et le taux de récidives aurait alors été de 54/280, soit 19 %.

– Ce n'est que si les 40 opérés perdus de vue ou décédés n'avaient pas fait de récidive à deux ans qu'il y aurait eu 14 récidives pour l'ensemble des 280 opérés, soit 5 %, évaluation la plus optimiste qu'il soit possible de faire.

La présentation des résultats par les auteurs de ce travail avait, au moins, le mérite d'indiquer avec précision le nombre de malades perdus de vue ou décédés avant la date de point. Cela permet à un lecteur critique, comme nous venons de le montrer, de nuancer l'interprétation des résultats.

Mesures

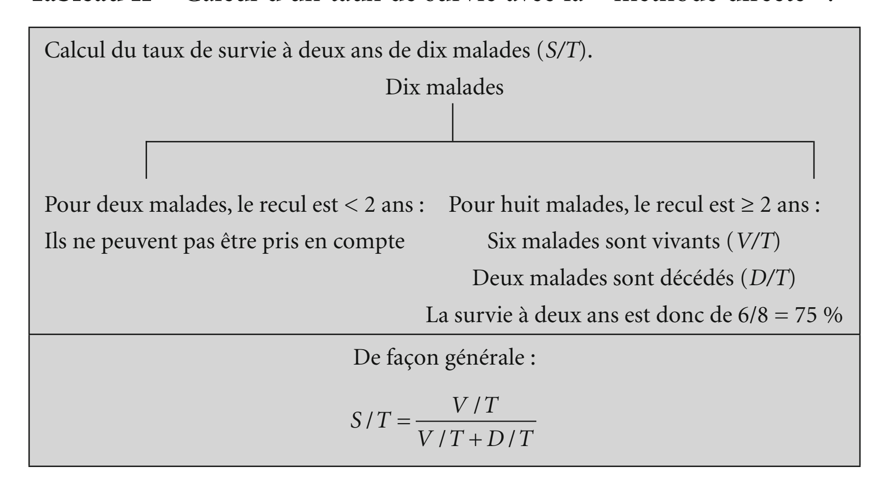
Des logiciels permettent de tracer les courbes de survie. Nous allons en décrire sur des exemples simples pour bien en faire comprendre le mécanisme, tout en sachant que la méthode couramment utilisée aujourd'hui est celle de Kaplan-Meier.

La méthode directe

La « méthode directe » n'est rappelée qu'à titre historique. En effet, elle ne prend en compte, dans les calculs, que les sujets pour lesquels le recul est suffisant. Les autres sujets, ainsi que les sujets perdus de vue, sont exclus de l'analyse.

Voici un exemple simple (tableau II). Si l'on veut évaluer le taux de survie à deux ans d'un groupe de dix malades pour lequel le recul pour deux malades est de moins de deux ans, ces derniers ne peuvent être pris en compte dans le taux du calcul du taux de survie à deux ans. Si, sur les huit autres malades, avec deux ans de recul, six sont vivants et deux sont décédés, le taux de survie à deux ans est de 6/8, soit 75 % (le calcul de l'intervalle de confiance à 95 % montre qu'il va de 45 % à 100 % !).

Tableau II – Calcul d'un taux de survie avec la « méthode directe ».



Calcul du taux de survie à deux ans de dix malades (*S/T*).

Dix malades

Pour deux malades, le recul est < 2 ans :	Pour huit malades, le recul est ≥ 2 ans :
Ils ne peuvent pas être pris en compte	Six malades sont vivants (*V/T*)
	Deux malades sont décédés (*D/T*)
	La survie à deux ans est donc de 6/8 = 75 %

De façon générale :

$$S/T = \frac{V/T}{V/T + D/T}$$

De la même façon, il est possible de mesurer le taux de survie à trois ans, à quatre ans, etc. Il s'agit d'estimations ponctuelles qui, en fait, peuvent être assimilées à l'évaluation de variables qualitatives. Elles ne doivent pas être reliées entre elles par une courbe. Du reste, cette méthode expose parfois au paradoxe d'un taux de survie à cinq ans supérieur à celui observé à quatre ans. C'est le cas si le nombre de malades, pour lesquels le recul est de cinq ans, est réduit et que, par le jeu du hasard, leur mortalité est moindre que celle observée dans le groupe qui a quatre ans de recul (fig. 1). Pour ces raisons, la « méthode directe » ne doit plus être utilisée pour mesurer une variable censurée, au profit de méthodes qui reposent sur des probabilités conditionnelles.

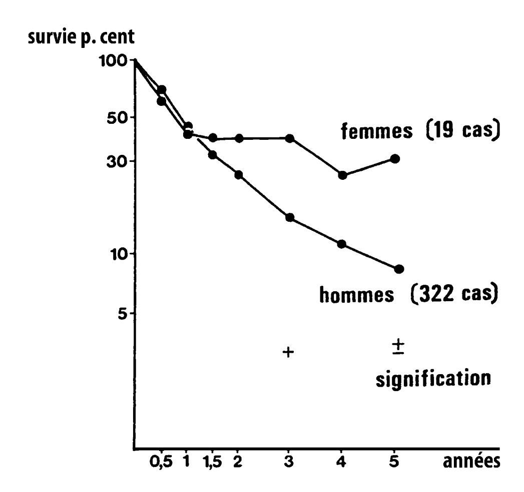


Fig. 1 – Expression d'une survie par la « méthode directe ». Dans cette étude, la survie chez la femme paraît augmenter entre 4 ans et 5 ans de recul (d'après Maillard *et al.* [1]).

Les probabilités conditionnelles

Leur principe est facile à comprendre en prenant l'exemple d'un jeu de cartes (tableau III).

Tableau III – Le principe des probabilités conditionnelles.

Dans un jeu de 52 cartes,
la probabilité de tirer l'as de pique est de 1/52.
En fait, cette probabilité peut se décomposer ainsi : c'est la probabilité de tirer un pique qui est de 1/4, multipliée par la probabilité de tirer l'as si l'on a tiré un pique qui est de 1/13.
1/4 multiplié par 1/13 est bien égal à 1/52.
Dans la survie,
la probabilité de survie à deux ans est égale à la probabilité de survie entre un et deux ans si l'on a survécu au moins un an.
Si la probabilité de survie à un an est de 87 %, et que la probabilité de survie entre un et deux ans est de 75 %, la probabilité de survie à deux ans est de 87 % × 75 %, soit 65 %.

La méthode de Kaplan-Meier [2]

Cette méthode est la méthode de choix pour l'analyse d'une variable censurée. Elle a l'avantage d'inclure dans l'analyse tous les sujets, quel que soit le recul d'observation. Elle constitue le moyen le plus précis et le plus clair d'exprimer la survenue d'événements qui dépendent du temps. Une « courbe » de Kaplan-Meier se présente, en fait, comme des marches d'escalier qui seraient de hauteurs et de largeurs différentes (fig. 2).

La construction d'une « courbe » de Kaplan-Meier, traduction graphique de la survie d'une population, permet de bien comprendre son principe et ce qu'elle représente.

Prenons comme exemple une population (ou un échantillon) fictif de neuf malades qui ont été opérés et chez lesquels on cherche à apprécier la survie. Il faut les classer par ordre croissant de recul (tableau IV).

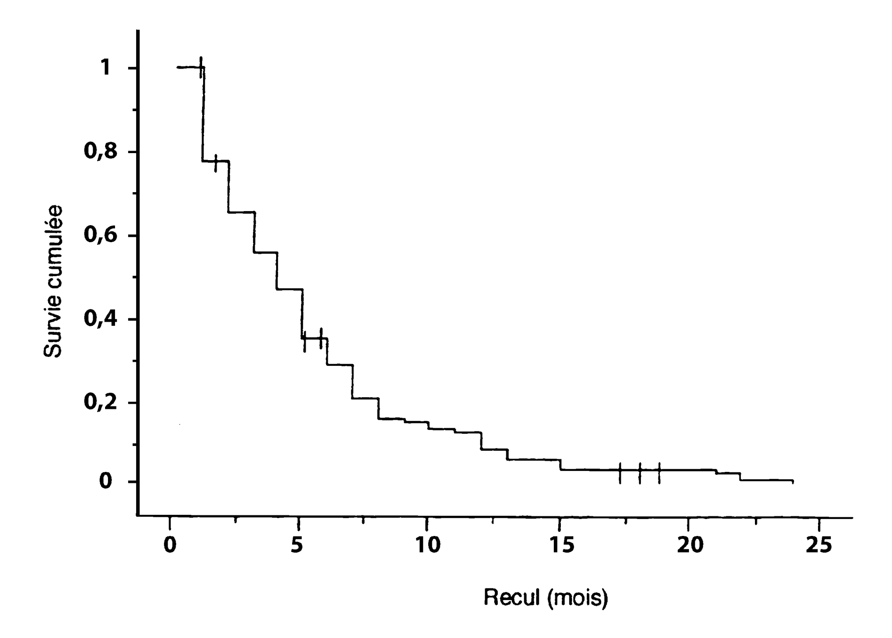
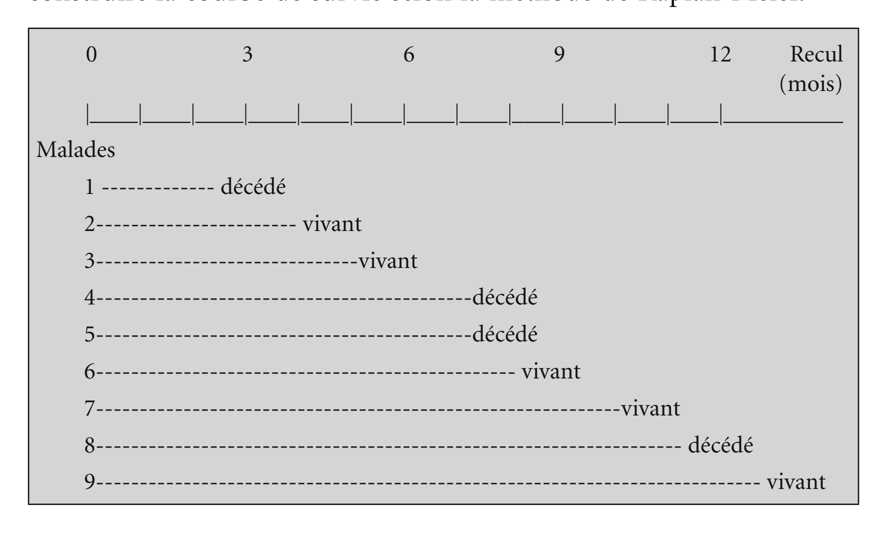


Fig. 2 – Expression d'une survie estimée par la méthode de Kaplan-Meier. Chaque marche de la courbe correspond à un décès (ou à plusieurs décès à la même date. Les petites barres verticales recoupant les horizontales représentent les malades exclus-vivants et le recul pour ces patients.

Tableau IV – Population fictive de neuf malades dont on va construire la courbe de survie selon la méthode de Kaplan-Meier.



Il convient ensuite de dresser le tableau suivant (tableau V).

Tableau V – Population fictive de neuf malades dont on va construire la courbe de survie selon la méthode de Kaplan-Meier.

Malades	Temps de partici-pation (mois)	a vivants en début d'intervalle	b décédé	c exclus-vivants	d probabilité de survie (a – b/a) condi-tionnelle	e survie cumulée (d.e)
1	2	9	1	0	8/9 = 0,89	0,89
2	4	8	0	1	1	0,89
3	5	7	0	1	1	0,89
4	7	6	1	0	1	0,89
5	7	6	1	0	4/6 = 0,68	0,60
6	8	4	0	1	1	0,60
7	10	3	0	1	1/2 = 0,50	0,30
8	11	2	1	0	1	0,30
9	12	1	0	1	1	0,30

Pour tracer la courbe de Kaplan-Meier (fig. 3), on part de l'abscisse temps zéro et de l'ordonnée 100 % de survie.

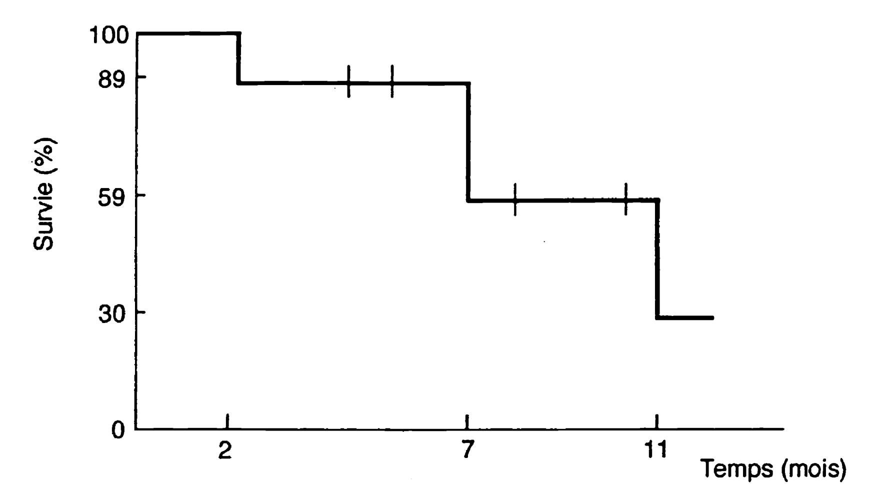


Fig. 3 – Tracé d'une courbe de Kaplan-Meier à partir de l'exemple théorique du texte.

Deux mois après l'opération, un malade est décédé. La « courbe » de survie, horizontale jusque-là, chute de 100 % à 89 % (il ne reste plus que huit survivants sur les neuf malades), ce qui se traduit par une première « marche d'escalier ».

Entre deux et sept mois de recul, deux patients sont exclus-vivants, l'un avec quatre mois de recul, l'autre avec cinq mois. Il est souhaitable de les faire figurer sur la « courbe » par une petite barre verticale sur l'horizontale de survie à 89 %, correspondant en abscisse au temps de participation à l'étude du malade.
À sept mois, deux malades décèdent. La survie cumulée chute alors à 60 % (4/6 × 8/9), ce qui détermine une nouvelle marche d'escalier horizontale correspondant en ordonnées à 60 %, etc.
Ainsi, la longueur des marches représente des intervalles pendant lesquels il n'y a pas eu d'événements que l'on cherche à estimer : décès ou récidive ou apparition d'une complication, etc. La descente d'une marche représente la chute du taux de survie lorsqu'un événement survient ; la survie, rappelons-le, dans le jargon statistique étant le délai qui sépare la date d'inclusion dans l'étude, de l'événement.

La méthode actuarielle

Dans la méthode actuarielle, les taux de survie, contrairement à l'estimation de Kaplan-Meier, sont évalués à intervalles réguliers, par exemple à un an deux ans, trois ans, etc. comme dans la méthode directe (fig. 4).

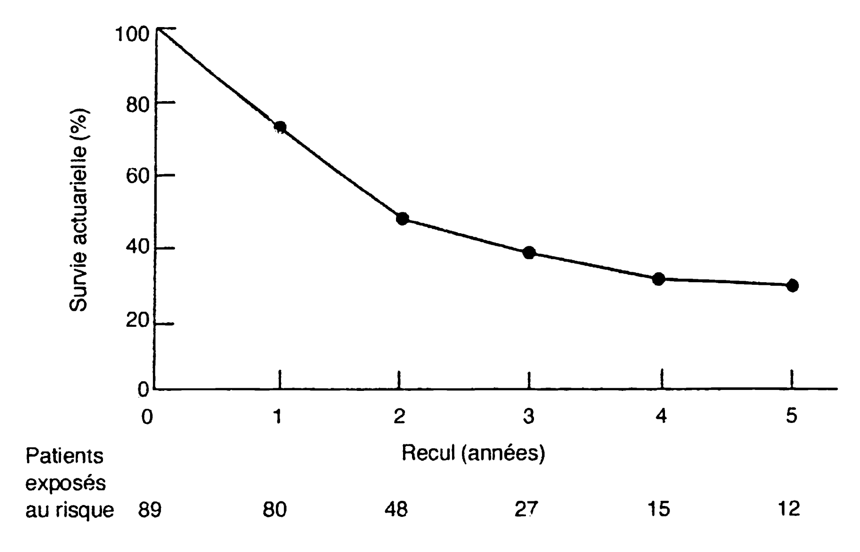
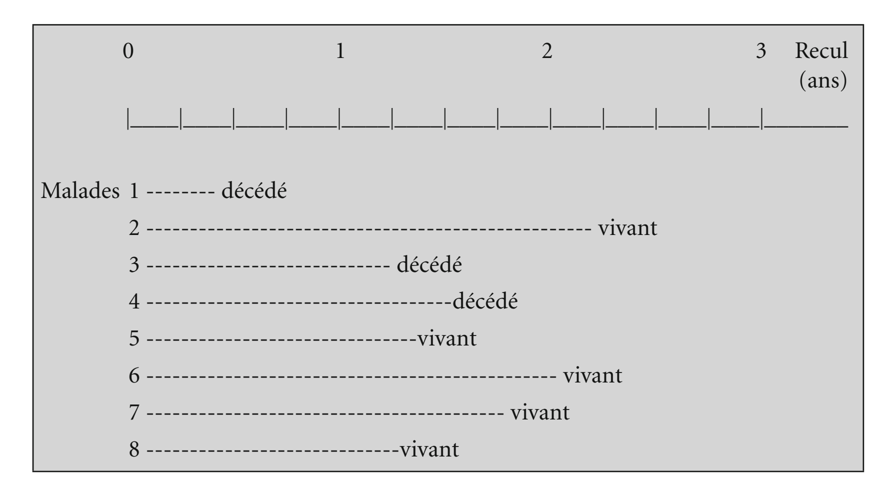


Fig. 4 – Courbe de survie actuarielle. Les taux sont calculés pour des reculs à intervalles prédéterminés (d'après Petrequin *et al.* [3]).

Nous allons construire une « courbe » actuarielle pour la comprendre en prenant comme exemple une population fictive de huit malades chez lesquels on a diagnostiqué un cancer et chez lesquels on cherche à apprécier la survie (tableau VI).

Tableau VI – Population fictive de huit malades
dont on va construire la courbe de survie actuarielle.



Pour construire une courbe actuarielle, il convient ensuite de dresser le tableau suivant (tableau VII).

Tableau VII – Construction d'une courbe actuarielle :
le calcul des survies.

Inter-valle	**a Vivants en début d'inter-valle**	**b Décédés dans l'inter-valle**	**c Vivants en fin d'inter-valle**	**d Exclus-vivants : a – (b + c)**	**e à risque dans l'intervalle b + c + (d/2)**	**f Survie en fin d'intervalle $\frac{e - b}{2}$**
1- 2 ans	7	2	1	7 – (2 + 1) = 4	2 + 1 + (4/2) = 3	3/5 = 60 %

Un malade décède dans la première année. Le taux de survie à un an est donc de 7/8, soit 89 %. Entre la première année et la seconde année, il y a quatre malades exclus-vivants (malades 5, 6, 7 et 8). Chez ces malades, les temps de participation à l'étude diffèrent cependant les uns des autres : il est assez court pour les malades 5 et 8. Il est plus long pour les malades 6 et 7. Cela permet d'estimer qu'en moyenne, un patient exclu-vivant sur deux est exposé au risque de décéder pendant l'intervalle ou, ce qui revint au même, qu'un patient exclu-vivant est exposé pendant la durée d'un demi-intervalle. De ce fait, le nombre de malades exposés au risque de décéder dans l'intervalle est considéré comme égal au nombre de malades vivants en début d'intervalle (n = 7), moins la moitié des exclus-vivants (4/2 = 2), soit 5. Cette prise en compte de la moitié des patients exclus-vivants à chaque intervalle

de temps est sous-tendue par l'hypothèse que les malades entrent régulièrement dans l'étude ou que les intervalles de temps sont suffisamment brefs. Cette hypothèse ne serait pas vérifiée si, par exemple, les intervalles étaient longs et que tous les patients exclus-vivants l'étaient, soit au début, soit en fin d'intervalle.
Le taux de survie en fin d'intervalle est égal au nombre de survivants en fin d'intervalle (n = 1) plus la moitié des exclus-vivants (4/2 = 2), soit 3 sur le nombre d'exposés au risque (n = 5), soit 3/5 = 60 %. La survie cumulée est ensuite calculée comme dans la courbe de Kaplan-Meier. La survie à un an étant de 89 %, la survie à deux ans est de 89 % × 60 %, soit 53 %, etc.

Quelques remarques doivent être faites :
1. Dans les deux méthodes, Kaplan-Meier et actuarielle, il est possible et souhaitable de calculer et de représenter l'intervalle de confiance à 95 % (habituellement par des pointillés comme sur la fig. 5). Parfois, on se contente d'indiquer l'intervalle de confiance en quelques points de la courbe, par exemple autour de la médiane de survie (correspondant à l'ordonnée 50 %) ou autour d'autres qui font l'objet de la discussion par exemple un taux de survie à deux ans ou à cinq ans, etc. (fig. 6).

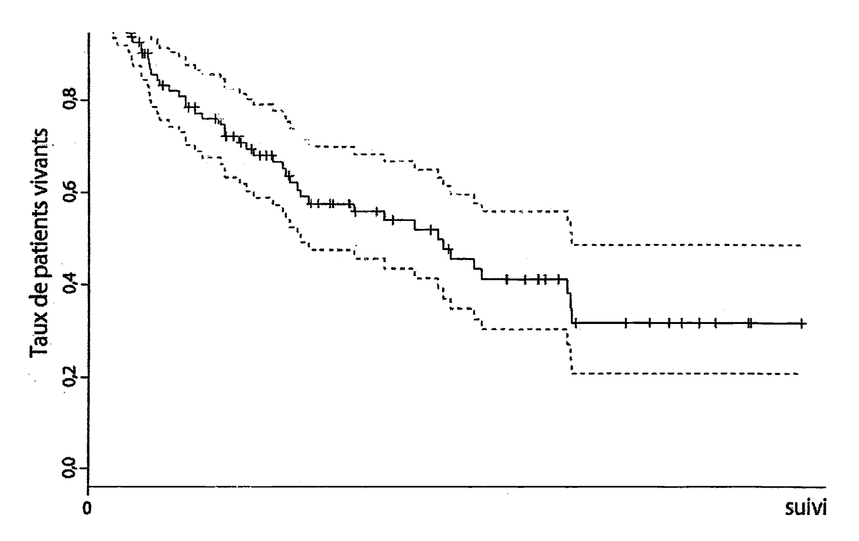


Fig. 5 – Courbe de Kaplan-Meier sur laquelle les exclus-vivants sont indiqués, les courbes en pointillés, représentant les limites des intervalles de confiance à 95 %.

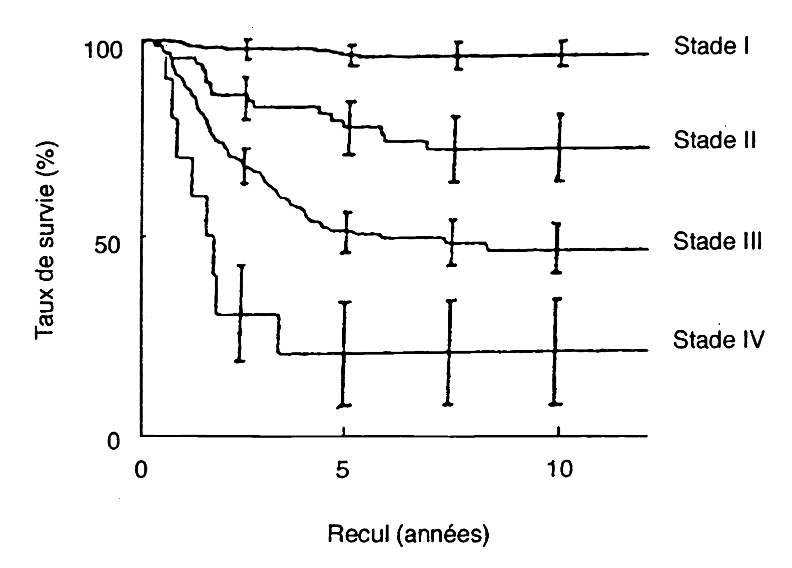


Fig. 6 – Courbes de Kaplan-Meier sur lesquelles les intervalles de confiance sont indiqués par des barres verticales (d'après Petrequin *et al.* [3]).

2. Il est encore souhaitable d'indiquer, sous l'axe des abscisses, le nombre de malades exposés au risque, par exemple à un an, deux ans, etc. (fig. 4).
3. Plus le taux de censure est important, plus les estimations seront imprécises.
4. Pour une même population, si l'on dressait et superposait la courbe de Kaplan-Meier et la courbe actuarielle, elles se confondraient d'autant plus que la population étudiée serait nombreuse et que, dans la courbe actuarielle, les intervalles de temps entre deux mesures seraient courts ; on peut presque dire qu'une courbe actuarielle est le lissage d'une courbe de Kaplan-Meier.

Références

1. Maillard JN, Huguier M, Conte-Marti J, Lortat-Jacob JL (1972) Pronostic éloigné de la résection pour cancer de l'œsophage. Nouv Presse Med 1: 2737-41
2. Kaplan EL, Meier P (1958) Non-parametric estimation from incomplete observation. Am Stat Ass J 53: 457-81
3. Pétrequin P, Huguier M, Lacaine F, Houry S (1997) Cancers de l'œsophage réséqués. Modèle prédictif de survie. Gastroenterol Clin Biol 21: 12-6

Les variables subjectives 5

À côté des variables qualitatives, quantitatives et censurées, on peut être amené à prendre en compte, dans une évaluation, des notions subjectives comme l'intensité d'une douleur, le confort ou bien, de façon plus générale, la qualité de vie, en particulier chez un malade qui a un cancer traité par exérèse chirurgicale, radiothérapie et chimiothérapie dont il subit, à côté des avantages attendus, des contreparties.
Pour évaluer ces variables subjectives, on cherche à les transformer en variables objectives, quantitatives. Un des exemples les plus anciens est l'indice de Karnofsky en cancérologie [1].

L'indice de Karnofsky en cancérologie		
Définition	**%**	**Critères**
Activité normale, travail normal	100	Pas de symptômes ni de signes de maladie
Aucun soin nécessaire	90	Activité normale, symptômes minimes
	80	Activité normale, quelques symptômes
Incapacité de travailler, vie à domicile possible	70	Soins personnels, mais incapacité d'activité professionnelle
Aide nécessaire pour la plupart des besoins	60	Assistance temporaire mais possibilité d'assumer la plupart de ses besoins personnels
	50	Nécessité d'une assistance et de soins médicaux fréquents
Perte d'autonomie	40	Handicapé, nécessité de soins spéciaux
Soins hospitaliers ou équivalents, maladie à progression rapide	30	Handicapé majeur, soins hospitaliers nécessaires
	20	Traitement médical de soins palliatifs
	10	Moribond
	0	Décès

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Moyens de mesure

Dans un phénomène douloureux, il est non seulement utile de savoir si un malade a mal ou non. Mais s'il a mal, il est encore plus utile de pouvoir apprécier l'intensité de la douleur. La transformation de données subjectives en variables quantitatives se fait généralement par des procédés d'auto-évaluation lorsque cela est possible, afin d'éliminer le biais que pourrait constituer l'appréciation par un tiers. Deux outils de mesure sont habituellement utilisés.

Avec **l'échelle visuelle analogique**, on mesure (au sens propre) l'intensité de la douleur par la distance entre le point situé entre le zéro qui représente l'absence de douleur et le trait indiqué par le sujet sur un segment de droite dont l'autre extrémité serait le « maximum » de douleur.

Échelle visuelle analogique

La ligne ci-dessous représente un « thermomètre de votre douleur.

Indiquez par un train vertical le niveau de votre douleur :

Pas de douleur *Maximum de douleur*

Il est encore possible de faire une quantification par une **cotation** (échelle de Likot).

Quantification par une cotation

Absente	ou pas du tout	0
Faible	ou un peu	1
Modérée	ou moyennement	2
Forte	ou beaucoup	3
Extrêmement forte	ou extrêmement	4

Donnez une valeur à votre douleur |_____|

Dans ce type de cotation, des études ont suggéré qu'il y avait peu de gain à proposer une cotation à plus de cinq niveaux couvrant uniformément la gamme des possibilités pour coter un phénomène subjectif. Il est évident que ces quantifications d'un phénomène subjectif laissent une part à l'interprétation. Elles constituent toutefois une approche intéressante pour mesurer ces phénomènes. Il sera important de

réaliser des mesures du niveau de base (*baseline* en anglais) afin de pouvoir apprécier, patient par patient, l'évolution de la variable. On s'affranchira ainsi de la subjectivité existant dans le niveau de base entre les patients.

D'autres mesures de phénomènes subjectifs sont plus complexes, comme l'évaluation de la **qualité de vie**. Ce type de mesure devient de plus en plus important avec le vieillissement de la population. En effet, on attache une importance plus grande à l'espérance de vie en bonne santé qu'à l'espérance de vie globale. Ainsi, en France, si l'espérance de vie était en 2006 de 85 ans chez la femme et de 77 ans chez l'homme, en bonne santé, elle n'était que de 73 ans (pour 75 ans au Japon et 70 ans aux États-Unis). Ces appréciations prennent en compte plusieurs variables subjectives qui peuvent être regroupées en rubriques (composantes psychologiques, comportementales, sociales, motrices, etc.). À chaque variable, il est possible d'affecter un coefficient en fonction de l'importance subjective qui lui est accordée ou qui est estimé à partir des résultats de vastes enquêtes [2].

Évaluation de qualité de vie

Composantes d'une évaluation [3]

1	Réactions émotionnelles
2	Énergie
3	Douleur
4	Sommeil
5	Isolement social

Avec le vieillissement des populations, des systèmes de mesure de l'autonomie fonctionnelle des personnes âgées (SMAF) sont de plus en plus utilisés. L'un d'entre eux comporte ainsi 29 items d'évaluation regroupés en cinq catégories : les activités de la vie quotidienne, la mobilité, la communication, les fonctions mentales et les activités de la vie domestique. Chaque item est coté sur une échelle de cinq degrés [4]. Ce type d'évaluation permet de rendre plus objectives les allocations de ressources aux établissements qui prennent en charge les personnes âgées.

Un autre outil de mesure est l'année de vie ajustée sur la qualité (QALY) [5]. Il concerne uniquement l'état de santé. C'est la somme des qualités de vie par année, la qualité de vie pouvant aller de 1 (état l'état de santé optimal) à 0 qui est le décès. Ainsi, 5 ans de qualité de vie moyenne de 0,5 donneront une qualité de vie-années de 2,5.

Références

1. Karnofski DA, Abelmann WH, Craver LF, Burchenal JH (1948) The use of nitrogen mustards in the palliative treatment of carcinoma. With particular reference to brochongenic carcinoma. Cancer 1: 634-56
2. Slim K, Bousquet J, Kiatkowsky F, *et al.* (1999) Première validation de la version française de l'index de qualité de vie pour les maladies digestives (GIOLI). Gastroenterol Clin Biol 23: 25-31
3. Hunt SM, McEwen J, McKenna SP (1985) Measuring health status: a new tool for clinicians and emidemiologists. J Roy Coll Gen Pract 35: 185-8
4. Gervais P, Hébert R, Jbaddi M, Toussignant M (2011) Implantation du système de mesure de l'autonomie fonctionnelle (SMAF) dans onze milieux d'hébergement et d'aide à domicile du secteur médico-social français : étude PISE-Dordogne. Revue de gériatrie 36: 631-44
5. Torrance GW (1987) Utility approach to measuring quality of life. J Chronic Dis 40: 593-600

Partie

Les comparaisons

2

Introduction

Les comparaisons sont une démarche habituelle dans toute activité biologique et médicale, aussi bien expérimentale que dans les sciences du vivant. En 1865, Claude Bernard écrivait déjà : « De tout cela je conclurai donc que l'observation et l'expérience comparative sont la seule base solide de la médecine expérimentale » [1].

En médecine, les comparaisons portent habituellement :

– sur différents examens radiologiques ou biologiques dans l'élaboration d'un diagnostic afin de choisir celui ou ceux qui paraissent les mieux à même de contribuer à ce diagnostic ;
– sur deux ou plusieurs traitements pour choisir celui qui est le plus efficace, le mieux toléré, etc. ;
– sur un ou plusieurs facteurs potentiels de risque dans l'estimation d'un pronostic ou en épidémiologie.

Ces comparaisons comportent des risques statistiques de conclusions erronées. On ne peut pas les éliminer, mais les limiter. Dans la lecture critique d'articles, leur connaissance permet de mieux interpréter les résultats.

Les comparaisons peuvent porter sur des pourcentages, des moyennes, des taux de survie, etc. c'est-à-dire sur les variables qualitatives, quantitatives, censurées, voire subjectives.

Exemples de comparaisons entre deux échantillons

Comparaisons portant sur des variables qualitatives*

– Évaluation diagnostique : comparaison des pourcentages de diagnostics de la résonance magnétique nucléaire et du Pet-Scan dans des métastases hépatiques.

– Évaluation thérapeutique : comparaison de deux antibiothérapies sur la stérilisation d'une infection urinaire.

– Évaluation pronostique en épidémiologie clinique : comparaison du rôle de l'existence ou non d'une insuffisance rénale dans le pronostic vital d'une pancréatite aiguë hémorragique.

– Évaluation étiologique en santé publique : comparaison de l'exposition ou non au tabac dans la survenue d'un cancer du poumon.

* Ici, les comparaisons ne sont pas explicitées par des quantités.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Comparaisons portant sur des variables quantitatives

– Évaluation diagnostique : comparaison des taux d'amylasémie selon l'existence ou l'absence de pancréatite*.

– Évaluation thérapeutique : comparaison des effets de deux médicaments antihypertenseurs sur la pression artérielle.

– Évaluation pronostique en épidémiologie clinique : après une embolie pulmonaire, comparaison de la valeur de la PO^2 dans le pronostic vital.

Comparaisons portant sur des variables censurées

– Évaluation thérapeutique : dans la maladie de Basedow, comparaison de deux antithyroïdiens de synthèse sur la survenue de récidive.

– Évaluation pronostique en épidémiologie clinique : comparaison du rôle pronostique de la présence ou de l'absence de métastases ganglionnaires d'un cancer sur la survie.

* Une hyperamylasémie n'est pas spécifique de pancréatite aiguë.

Toute comparaison est exposée à des biais qui peuvent fausser l'interprétation des résultats. Il convient de prévenir au maximum ces biais par un plan expérimental adapté.

Le biais le plus commun est de faire porter les comparaisons sur des sous-groupes qui ne sont pas similaires. Pour limiter ce risque, il faut, au sein de l'ensemble de la population incluse dans l'étude, faire un tirage au sort pour déterminer deux (ou plusieurs) sous-groupes sur lesquels vont porter les comparaisons. C'est la randomisation.

En fin d'étude, la comparaison des résultats, même entre des sous-groupes qui seraient parfaitement similaires en dehors de la variable d'intérêt, est exposée à plusieurs risques d'interprétation. Ces risques sont les suivants :

- *Le premier risque*, à la vue d'une différence (A > B), est de conclure que A est supérieur à B alors que la différence observée est le fait du hasard et qu'il n'y a pas de différence réelle entre A et B. C'est le risque de première espèce, encore appelé *risque* α.
- *Le second risque* est celui d'une absence de différence significative (A = B) alors même que A est différent de B. C'est le risque de seconde espèce, encore appelé *risque* β.
- *Le risque de troisième espèce* est de conclure à tort que A > B ou A < B alors que c'est l'inverse. Ce *risque*, encore appelé γ, est largement plus faible que les précédents. Il est donc généralement négligeable. Ses conséquences seraient toutefois plus graves.

À côté de ces risques, il faut bien différencier les études dont le but est de chercher une différence, en général pour montrer la supériorité d'un examen ou d'un traitement sur un autre, études qui sont une démarche habituelle, notamment en recherche clinique, des études dont le but est de prouver une absence de différence, comme dans les essais d'équivalence ou de non-infériorité, études qui sont de plus en plus fréquentes dans tous les domaines de la recherche biomédicale.

Risques d'erreurs dans toute comparaison entre deux échantillons (A et B)

Ce qui a été déduit à partir des résultats observés :	La réalité : A > B	A = B	A < B
A > B	Correct	Erreur de 1re espèce	Erreur de 3e espèce*
A = B	Erreur de 2e espèce	Correct	Erreur de 2e espèce
A < B	Erreur de 3e espèce*	Erreur de 1re espèce	Correct

* Ce risque est faible et généralement négligeable.

Les méta-analyses seront étudiées avec le chapitre consacré à la thérapeutique parce que c'est essentiellement dans ce domaine qu'elles sont réalisées.

Référence

1. Bernard C (1865) Introduction à la médecine expérimentale. Baillière, Paris, p 342

Protocole médical d'un essai randomisé 1

Tout travail cherchant à comparer entre eux deux (ou plusieurs) « outils » diagnostiques ou traitements ou facteurs de risque, s'expose à un premier biais : que les sous-groupes sur lesquels porte la comparaison ne soient pas similaires. Ce genre de comparaisons a donné lieu à bien des erreurs de jugement dont les conséquences ont été lourdes. Par exemple, elles ont fait croire, dans le traitement des pancréatites aiguës, que les inhibiteurs de la trypsine diminuaient la mortalité, alors qu'il n'en est rien. Des malades ont ainsi été traités inutilement avec un médicament qui était onéreux puisqu'une année, il avait représenté le second poste de dépenses médicamenteuses de l'Assistance publique-Hôpitaux de Paris.

Dans toute comparaison sur la valeur d'un nouvel examen par rapport à un autre ou d'un traitement par rapport à un traitement de référence, il convient de faire porter les comparaisons sur des sous-groupes similaires. C'est, comme il a été indiqué dans l'introduction de cette partie, l'objectif des études ou essais randomisés.

Ils consistent, au sein de la population incluse dans l'étude (et clairement définie au départ), à déterminer les deux (ou plusieurs) sous-groupes sur lesquels porte la comparaison, par un tirage au sort. C'est ce qui offre le plus de chances que ces sous-groupes soient similaires. Encore peut-on s'en assurer *a posteriori*. En fin d'étude, il est déconseillé de faire des tests statistiques sur les caractéristiques à l'inclusion des groupes que l'on compare (règle CONSORT). En effet, d'éventuelles différences significatives entre groupes n'auront un retentissement que si elles ont à une pertinence médicale. Inversement, des différences, même non significatives peuvent biaiser les résultats de l'étude.

Il faut encore que la randomisation soit faite dans des conditions rigoureuses. En effet, le tirage au sort pour un essai randomisé est moins simple qu'on pourrait le croire. Ces essais randomisés sont des entreprises lourdes.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Les buts de ce chapitre sont de montrer les contraintes d'une randomisation correcte, ses limites pour des raisons techniques ou éthiques. En tant que lecteur, lorsque l'on veut se faire une opinion sur une comparaison entre deux « outils » diagnostiques ou entre deux traitements, il faut de façon préférentielle : 1) lire les résultats des essais randomisés et 2) être capable d'avoir une opinion critique sur leur méthodologie et l'interprétation de leurs résultats.

Nous prendrons surtout comme exemple, des études sur un traitement.

Le préalable à tout essai randomisé

Il n'est moralement licite de faire un essai thérapeutique randomisé pour chercher la supériorité d'un traitement par rapport à un autre que si l'on accepte une double hypothèse presque paradoxale : 1) on cherche à montrer qu'un traitement est meilleur qu'un autre et 2) on doute de cette éventualité.

Si l'objectif de l'étude est de montrer l'équivalence entre deux traitements, la double hypothèse devient la suivante : 1) les traitements sont probablement équivalents entre eux, mais 2) on doute de cette éventualité que l'on cherche à prouver.

Ces propositions seront tranchées grâce au critère de jugement principal de l'étude.

Inclusion des sujets dans l'étude

Dans un essai randomisé, une première condition d'inclusion est que tout malade inclus dans l'étude doive pouvoir recevoir l'un ou l'autre des deux traitements que l'on cherche à comparer. C'est ce qui est appelé la **clause d'ambivalence**. Par exemple, dans des essais randomisés comparant dans le cardiospasme le traitement endoscopique par des dilatations pneumatiques et le traitement chirurgical par myotomie extramuqueuse (opération de Heller), il faut que tous les malades inclus dans l'étude puissent, éventuellement, être opérés et acceptent cette éventualité. S'il y a une contre-indication opératoire, cela doit représenter un critère d'exclusion. Les bonnes études doivent préciser le nombre de malades qui ont été exclus, ainsi que les causes de l'exclusion. Dans notre exemple, ce peut être une contre-indication opératoire, mais aussi le fait des malades, qui avaient accepté l'idée de se faire opérer, l'ont ensuite refusée après le tirage au sort les désignant dans le groupe chirurgical. Pour cette raison, et de façon générale, il est souhaitable que l'intervalle entre le tirage au sort et le début du

traitement soit aussi bref que possible. Dans un essai chirurgical comparant deux techniques, le tirage au sort doit être fait lors de l'intervention, lorsque le chirurgien s'est assuré que la clause d'ambivalence était respectée. La panseuse, par exemple, indique alors au chirurgien, après ouverture d'une enveloppe, celle des deux techniques qu'il doit réaliser.

La connaissance du nombre de malades exclus de l'étude et les raisons d'exclusion permettent de se faire une opinion sur le champ d'application des résultats de l'étude. En effet, pour fondamentaux que soient les essais randomisés, l'expérience montre que les critères d'inclusion aboutissent à ce qu'elles intéressent des populations assez sélectionnées. En théorie, leurs résultats ne peuvent donc être utilisés, s'ils sont positifs, que pour des malades, eux aussi assez sélectionnés sur les mêmes critères que ceux sur lesquels l'étude avait porté.

Précautions concernant les traitements que l'on cherche à évaluer

Si l'on cherche à évaluer un traitement médical par rapport à l'absence de traitement, l'idéal est que les malades du groupe témoin reçoivent un placebo, c'est-à-dire un comprimé, une gélule, ou une potion d'aspect similaire au comprimé, à la gélule ou à la potion du principe actif, mais qui ne le contient pas. S'il s'agit d'un mode d'administration intraveineuse, cela est plus difficile pour des raisons d'éthique.

Néanmoins, une des critiques faite aujourd'hui à la Commission de transparence qui dépend de la Haute autorité de santé est de fonder trop souvent ses décisions sur le service médical rendu par un nouveau médicament en le comparant à un placebo. Il serait beaucoup plus utile et souhaitable de le comparer avec un médicament existant, c'est-à-dire de fonder la décision non sur l'existence d'un service médical rendu, mais sur une amélioration d'un tel service par rapport à un traitement préexistant de référence. Les méta-analyses en réseau permettent notamment d'estimer les différences entre traitements lorsque ceux-ci n'ont pas été comparés directement entre eux, mais l'un et l'autre avec un même contrôle (ou placebo).

Afin de limiter le risque de biais psychologiques, il est encore souhaitable que le malade ignore s'il reçoit le principe actif ou le placebo (ou soit un traitement A, soit un traitement B). On parle alors d'étude en **simple insu** (*blind* en anglais). Lorsque le médecin ou l'infirmier qui administre le médicament ignore, lui aussi, le contenu réel du traitement alloué, on parle d'essai en **double insu**. C'est alors un tiers

qui, seul, connaît ce que reçoit le malade jusqu'à la fin de l'étude où l'anonymat du produit administré est dévoilé. Le but est de garantir le maximum d'objectivité dans le recueil des résultats. La possibilité des biais dans ce domaine n'est pas purement théorique. Ainsi, un essai randomisé comparant l'effet d'un patch de nicotine avec un patch de placebo pour faciliter le sevrage du tabagisme [1] a prouvé que ces précautions n'étaient pas superflues. En effet, les résultats ont montré, d'une part que 4 % des fumeurs qui avaient reçu le placebo ont dit avoir ressenti des contreparties telles qu'ils ont arrêté de se mettre le patch, d'autre part que 16 % d'entre eux se sont arrêtés de fumer et ont attribué cet arrêt au patch qui s'est avéré être le placebo.

Dans les essais randomisés comportant une intervention chirurgicale, le double insu, bien entendu, n'est pas possible, le chirurgien sachant forcément ce qu'il a réalisé comme intervention ! Dans ce cas, il est possible de faire évaluer les résultats par un tiers. Dans les essais sur une technique chirurgicale ou radiologique, il existe un autre biais : si l'on cherche à comparer une technique chirurgicale de référence à une nouvelle technique, le chirurgien ou le radiologue va comparer une technique qu'il connaît bien et à la réalisation de laquelle il est entraîné, à une technique dont il a moins l'expérience puisqu'elle est nouvelle. Pour limiter le risque créé par ce biais, il est souhaitable que le chirurgien, par exemple, commence par opérer des malades avec la nouvelle technique pour avoir un bon entraînement dans la réalisation de la nouvelle technique, c'est la courbe d'apprentissage, avant de commencer à inclure des malades dans l'essai [2]. Une valeur seuil correspond au nombre d'interventions nécessaires pour acquérir une expérience [3]. Plus cette valeur est élevée, plus l'intervention est jugée difficile. Par exemple, en chirurgie colorectale par cœlioscopie, une étude multifactorielle a suggéré que la valeur seuil de la courbe d'apprentissage était de 40 interventions [4]. Dans les fundoplicatures par cœlioscopie, la valeur seuil a été estimée à 20 [5], ce qui suggère que l'apprentissage de cette intervention est moins difficile que celui de la précédente.

Les comparaisons portant sur le rôle d'un examen complémentaire, qu'il s'agisse d'une exploration radiologique, mais aussi biologique ou isotopique, peuvent et doivent, elles aussi, faire l'objet d'essais randomisés.

Les examens complémentaires, s'ils ne sont pas invasifs, c'est-à-dire sans risque pour le malade, permettent cependant de les comparer chez un même malade, en les réalisant chez chacun des malades inclus dans l'étude. Chaque malade est alors son propre témoin, ce qui permet de réduire notablement les biais dus aux caractéristiques individuelles des patients. La similitude entre les groupes

comparés est alors maximale, puisqu'ils sont constitués de personnes aux mêmes caractéristiques. Bien entendu, ces études nécessitent le consentement éclairé des patients. Il convient encore que le temps qui sépare les deux examens l'un de l'autre soit réduit au maximum dans le cas où il s'agit d'examen morphologique, radiologique par exemple.

Les critères de jugement

Dans un essai randomisé, il convient toujours, lors de l'élaboration de l'étude, de décider, parmi les critères de jugement pertinents, celui qui est le plus important. Ce sera le **critère de jugement principal**. Les autres sont appelés critères de jugement secondaires. C'est le critère de jugement principal qui permettra de conclure à l'efficacité plus grande d'un traitement par rapport à un placebo ou à un autre traitement. C'est aussi lui qui va permettre de déterminer le nombre de sujets qu'il est souhaitable d'inclure dans l'étude pour limiter le risque de deuxième espèce. Plus le nombre de critères de jugement secondaires est élevé, plus on augmente le risque d'observer une différence significative pour l'un deux qui, en fait, est due au hasard. De plus, la décision que l'on pourra prendre au vu des résultats de l'étude sera d'autant plus compliquée que les critères de jugement sont nombreux, qu'ils soient d'ordre médical comme les contreparties des traitements que l'on compare ou bien d'ordre économique. Ainsi, dans le traitement d'un cancer, si une chimiothérapie s'avère plus efficace qu'une autre en termes de durée de survie, mais que le gain de survie est relativement limité dans le temps et se fait au prix de contreparties qui altèrent la qualité de cette survie, la décision de traiter ne sera pas forcément prise en prescrivant la chimiothérapie la plus efficace en termes de durée de survie.

La détermination des critères de jugement implique donc une réflexion qui n'est pas toujours aisée. Ne pas prendre en compte un critère de jugement secondaire au sens statistique, mais important comme la contrepartie sévère d'un traitement, serait préjudiciable. Mais multiplier les critères de jugement complique la réalisation de l'étude, augmente comme nous le verrons le risque global de première espèce, et surtout l'interprétation décisionnelle de ses résultats.

En tout état de cause, il est particulièrement important dans un souci d'objectivité, que le médecin qui évalue les résultats d'un essai randomisé le fasse sur des critères de jugement aussi précis que possible et ignore le traitement qui a été administré (ou le placebo).

Les liens entre ces différentes données

Dans les essais randomisés, il y a presque toujours des interactions entre les trois données précédentes. Ainsi, lorsque le critère de jugement principal est la survie à 5 ans, il ne faut inclure dans l'étude que des malades qu'il sera possible de suivre jusqu'à leur décès dans les 5 années suivantes et, *a contrario*, exclure par exemple des malades vivant dans un pays étranger et pour lesquels on n'est pas certain de connaître cinq ans après s'ils sont toujours en vie ou s'ils sont décédés. Cette clause doit ainsi figurer dans le protocole d'inclusions.
Si l'essai porte sur une polychimiothérapie comprenant un médicament cardiotoxique, les critères d'inclusion doivent comporter un examen cardiovasculaire normal. Inversement, des antécédents ou des anomalies cardiaques pourront être des critères d'exclusion, etc.

Références

1. Jorenby DE, Leibshow SJ, Nides MA, *et al.* (1999) A controlled study of sustained release bupropion, a nicotine patch, or both for smoking cessation. New Engl J Med 340: 685-62
2. Bells PRF (1997) Surgical research and randomized trials. Br J Surg 84: 737-8
3. www.maaw.info/LearningcurveSummary.htm
4. Bennett CL, Stryker SJ, Ferreira R, *et al.* (1997) The learning curve for laparoscopic colorectal surgery. Arch Surg 132: 41-5
5. Watson DI, Baigrie RJn, Jamieson GG (1996) A learning curve for laparoscopic fundoplication. Definable, avoidable, or waste of time? Ann Surg 224: 198-203

Protocole statistique d'un essai randomisé 2

Ce protocole comporte, dans l'ordre chronologique de son élaboration :

– le calcul des effectifs pour limiter le risque de deuxième espèce ;
– les modalités de la randomisation ;
– les méthodes d'analyse des résultats.

Dans un souci didactique, nous parlerons d'abord du tirage au sort, puis des risques statistiques et enfin des problèmes de l'analyse des résultats.

Le tirage au sort

Au sein de la population incluse dans l'étude, le tirage au sort est le moyen qui offre les meilleures garanties que les sous-groupes qu'il détermine soient similaires, sauf en ce qui concerne le traitement alloué. Les différences observées en fin d'étude pourront ainsi être clairement attribuables aux traitements, c'est-à-dire à ce que l'on cherche à évaluer. C'est bien ce qui fait la spécificité et l'intérêt des essais randomisés par rapport aux comparaisons avec recueil rétrospectif des données. En effet, dans ces études, il ne sera pas possible de garantir que les deux sous-groupes comparés étaient similaires au départ.
Le tirage au sort est cependant moins simple qu'il pourrait paraître.

Les tirages au sort les plus simples ont des défauts

Un moyen facile de réaliser un tirage au sort serait d'utiliser une pièce de monnaie, le côté face indiquant l'allocation du traitement A et le côté pile celle du traitement B. Dans le même ordre d'idée, on pourrait utiliser le chiffre pair ou impair du jour de naissance du

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

malade ou de son jour d'hospitalisation, de sa carte Vitale, etc. Cette façon de procéder a deux inconvénients.

Le premier est que le médecin connaît d'emblée le traitement qui devra être administré au malade, avant même de s'être assuré que les conditions d'inclusion dans l'essai soient bien remplies. Cela risque de l'influencer en n'incluant pas certains malades parce qu'il estime qu'il est peut-être préférable qu'ils ne reçoivent pas le traitement indiqué par le chiffre pair ou impair. Il est, en effet, indispensable qu'un malade : 1) réponde d'abord aux critères d'inclusion (et d'exclusion), c'est-à-dire soit « éligible » ; 2) signe un consentement libre et éclairé comme nous le verrons ; 3) avant que le tirage au sort désigne le traitement qu'il doit recevoir. C'est ce que l'on appelle la **clause d'ignorance** au moment du tirage au sort.

Le second inconvénient des tirages au sort « simples » est qu'avec un peu de malchance, les deux groupes de malades soient quantitativement déséquilibrés. Ce risque est particulièrement élevé si l'on inclut dans l'étude un petit nombre de malades, par exemple dans l'un des centres d'un essai multicentrique. Ainsi, une étude avait été faite sur l'intérêt éventuel de l'atropine dans le traitement des pancréatites aiguës. Cinquante et un malades avaient été inclus dans l'étude. Le tirage au sort avait été fait, écrivaient les auteurs, sur la base « de nombres historiques » (il devait s'agir de nombres comme la date de naissance que nous avons évoquée). La malchance a fait que 19 malades ont reçu de l'atropine et 32 ont fait partie du groupe témoin. Plus le nombre de malades inclus dans l'étude est faible, plus le risque de déséquilibre quantitatif des sous-groupes, déterminés par un tirage au sort « simple », est important, ce qui diminue ce que l'on appelle la puissance des tests statistiques (cf. p. 93).

Tables de nombre au hasard (encore appelés nombres aléatoires) et de permutation de nombres au hasard

Pour limiter les inconvénients et les risques des méthodes précédentes, on utilise des tables de nombres au hasard. Ces nombres sont fournis par des ordinateurs[1] et figurent dans des tables qui se présentent sous forme d'une série de chiffre (tableau I). En les prenant successivement, il suffit de décider que les malades qui ont un nombre pair auront le traitement A et un nombre impair le traitement B. Pour respecter la clause d'ignorance, ces indications thérapeutiques sont mises chacune sous enveloppe numérotée. À l'inclusion du premier

1 Ces tables sont disponibles sur http://perso.orange.fr/jpq/proba/tablealea/index.htm

Tableau I – Table de nombres au hasard.

26099	65801	69870	84446	58248	21282	56938	54729	67757
71874	61692	80001	21430	02305	59741	34262	15157	27545
08774	29689	42245	51903	69179	96682	91819	60812	47631
37294	92028	56850	83380	05912	29830	37612	15593	73198
33912	37996	78967	57201	66916	73998	54289	07147	84313
63610	61475	26980	23804	54972	72068	19403	53756	04281
01570	41701	30382	54647	06077	29354	95704	75928	21811
24159	77787	38973	82178	46802	90245	01805	23906	96559
92834	52941	88301	22127	23459	40229	74678	21859	98645
16178	60063	59284	16279	48003	44434	08623	32752	40472
81808	32980	80660	98391	62243	19678	39551	18398	36918
28628	82072	04854	52809	86608	68017	11120	28638	72850
62249	65757	12273	91261	96983	15082	83851	77682	81728
84541	99891	01585	96711	29712	02877	70955	59693	26838
89052	39061	99811	69831	47234	93263	47386	17462	18874
13407	62899	78937	90525	25033	56358	78209	47008	72488
50230	63237	94083	93634	71652	02656	57532	60307	91619
84980	62458	09703	78397	66179	46982	67619	39254	90763
22116	33646	17545	31321	65772	86506	09811	82848	92211
68645	15068	56898	87021	40115	27524	42221	88293	67592
26518	39122	96561	56004	50260	68648	85596	83879	90941
36493	41666	27871	71329	69212	57932	65281	57233	07732
77402	12994	59892	85581	70823	53338	34405	67080	16568
83679	97154	40341	84741	08967	73268	94952	59008	95774
71802	39356	02981	89107	79788	51330	37129	31898	34011
57494	72484	22676	44311	15356	05348	03582	66183	68392
73364	38416	93128	10297	11419	82937	84389	88273	96010
14499	83965	75403	18002	45068	45257	18085	92625	60911
40747	03084	07734	88940	88722	85717	73810	79866	84853
42237	59122	92855	62097	81276	06318	81607	00565	56626
95307	65668	21280	75514	68955	57328	74675	67958	37864
79748	67309	46843	19734	45248	20343	77530	06735	53622
00586	33144	36553	57446	66156	31637	15924	71923	73089
85120	18976	42639	67159	86473	79129	02003	08708	65678
35493	36645	23427	12223	67361	19073	39770	13548	64994

malade, on ouvre la première enveloppe, du second la seconde enveloppe, et ainsi de suite. En fait, actuellement la randomisation est centralisée par Internet.

Il est souhaitable de faire le tirage au sort qui détermine le traitement alloué, juste avant l'institution du traitement. En effet, si ce délai est trop long, le malade risque, comme nous l'avons déjà évoqué, de changer d'avis et de refuser le traitement qu'il avait accepté initialement. L'inconvénient de ces tables de nombre au hasard, comme le montre le tableau I, est que sur les 30 premiers nombres au hasard de la première colonne, il y a 19 nombres pairs et 11 nombres impairs, déséquilibre que l'on cherche à éviter, comme nous l'avons vu. De plus, si le rythme d'inclusion est lent, on s'expose à l'inconvénient des comparaisons historiques. Par exemple, si l'on inclut 22 malades en deux ans, sur les 11 premiers malades inclus en 2010, deux auraient reçu le traitement B et sur les 11 derniers malades inclus en 2011, cinq. Or, des progrès autres que les traitements comparés ont pu intervenir entre les deux périodes.

Pour cette raison, on utilise plutôt des tables de permutation de nombres au hasard (tableau II). Dans ces tables, chaque groupe vertical de neuf chiffres a un dernier nombre qui va de 1 à 9. On peut alors décider, par exemple, que lorsque ce dernier chiffre est 1 ou 2, les malades recevront le traitement A et lorsque ce chiffre est 3 ou 4, ils recevront le traitement B. L'avantage est que, une fois tous les quatre malades inclus dans l'étude, il y en aura autant (deux) qui auront reçu le traitement A que le traitement B. On parle alors de randomisation par « blocs de 4 ». Dans ce type de randomisation, si l'essai n'est pas en double aveugle, le médecin saura, après inclusion des trois premiers malades, ce que le quatrième devra recevoir comme traitement, dérogeant ainsi au principe d'ignorance. Pour cette raison, on peut choisir de varier de façon aléatoire la taille des blocs de randomisation, par exemple 4, 6, 8. La taille des groupes de permutation peut être choisie en fonction du rythme d'inclusion des malades dans l'étude : petite taille si ce rythme est lent et inversement. Il est encore souhaitable que celui qui prépare, à l'aide de la table de permutation de nombre au hasard, les enveloppes (ou le programme informatique) dans lesquelles est indiqué le traitement qui sera alloué ne soit pas celui ou ceux qui administrent ces traitements.

Tableau II – Table de permutation de nombres au hasard.

55671	43373	87463	97494	92288	27935	83194
41282	71129	95782	89366	17724	48573	37456
93329	88845	24616	36778	74471	73286	61222

79743	55292	16535	78519	51913	65149	29878
16965	69436	43929	51823	83332	89612	45769
64436	24681	79341	62642	29859	92428	96981
87817	12568	31298	44i87	65167	54351	14317
32194	36757	68877	25951	38546	36794	52545
28558	97914	52154	13235	46695	11867	78633
74615	92229	28173	24219	24831	26548	84942
93832	11198	94954	88886	77546	53276	93821
16347	65845	61719	52563	85755	69981	36797
68284	48786	57545	96758	. 59977	85335	69469
41478	23934	42236	47425	63369	17854	45214
29193	79662	16461	79974	18418	92793	18355
55551	37477	85892	15132	96284	38119	57133
82929	86553	79688	31697	41693	44662	72688
37766	54311	33327	63341	32122	71427	21576
97755	99938	98617	58612	19833	31773	76655
38172	62716	41342	36243	26128	88627	89747
43427	73172	15486	62161	78517	59136	31231
59283	37589	29171	23834	35999	72341	57178
16511	56441	73723	47388	93256	66959	98912
62836	84625	52268	91756	47464	17464	12886
24964	18354	36594	85979	81681	45595	24594
85699	25267	87839	19425	64745	23282	63323
71348	41893	64955	74597	52372	94818	45469
74987	97171	92387	78535	51649	78618	29734
56112	64614	59128	24687	73761	51741	93477
49356	11848	35493	36123	26877	45385	85951
33228	52322	73869	41861	19236	39577	12812
21494	46283	27651	57312	98413	63129	61588
97545	39799	14234	69744	32522	84263	56363
62639	88555	86772	93458	87994	92494	48129
85871	23937	41515	85976	45358	16852	34645
18763	75466	68946	12299	64185	27936	77296
84686	21997	22189	51924	52628	16883	81941
99458	44878	87597	36477	38536	44677	66878
66311	68319	75755	65185	24382	51436	49786
73772	73622	38946	47269	79741	38265	35314
28934	15551	54364	78753	95865	82792	53435
37269	86463	41821	19648	47213	63551	22699
51845	99184	19432	82896	63499	27124	98262
45527	32736	93218	93512	16977	95918	77157
12193	57245	66673	24331	81154	79349	14523

Problèmes particuliers

La stratification

Un tirage au sort insuffisamment pensé ne met pas toujours à l'abri d'une différence entre les sous-groupes que l'on souhaite comparer, ce qui complique singulièrement l'interprétation des résultats. Cette différence, bien que due au hasard, peut être gênante lors de cette interprétation si elle porte sur une caractéristique associée à la valeur du critère principal. C'est ce qui a eu la malchance de se produire dans un important essai randomisé sur les cancers oto-rhyno-laryngologiques (ORL). L'idée de cet essai reposait sur des études antérieures qui suggéraient que la radiothérapie était plus efficace sur des tissus bien oxygénés que l'inverse. Sur la base de cette donnée, un essai randomisé a été réalisé pour comparer, chez des malades qui avaient un cancer ORL, la radiothérapie simple à la radiothérapie associée à l'administration d'oxygène hyperbare [1]. Le critère de jugement était la survie. Les résultats ont été similaires dans les deux groupes. Mais, une fois l'étude terminée, les auteurs se sont aperçus que, malgré le tirage au sort, les deux sous-groupes de malades n'étaient pas similaires. Le hasard a fait que les malades qui avaient eu de l'oxygène hyperbare avaient plus souvent des métastases ganglionnaires que les malades de l'autre groupe. Or les métastases ganglionnaires sont un facteur de mauvais pronostic. Ainsi, l'absence de meilleurs résultats dans le groupe oxygène hyperbare pouvait être due au fait que l'oxygène n'améliorait pas l'efficacité de la radiothérapie. Mais l'autre hypothèse était que cette amélioration avait été masquée par le déséquilibre induit, par hasard, par la présence différentielle des métastases ganglionnaires.
Afin de limiter le risque de tels déboires dans un essai randomisé, lorsque l'on sait qu'il existe un facteur de pronostic reconnu comme étant très important (dans notre exemple de cancer ORL, des métastases ganglionnaires), il est prudent de faire une stratification. Cela consisterait à réaliser un tirage au sort différent pour les malades qui n'ont pas de métastases ganglionnaires et ceux qui en présentent. Cette façon de procéder garantit qu'à la fin de l'inclusion et de l'essai, la proportion de malades qui ont des métastases ganglionnaires sera la même dans les deux sous-groupes que l'on cherche à comparer.
Sur le plan mathématique, la stratification augmente ainsi la puissance des tests d'inférence statistique en diminuant la variance du critère de jugement. En contrepartie, les modalités du tirage au sort deviennent un peu plus complexes et donc plus sujettes à des causes d'erreur de la part des investigateurs. En pratique, il n'est souhaitable de recourir à la stratification que lorsqu'elle semble vraiment justifiée et de ne pas dépasser un, voire deux niveaux de stratification.

Randomisation et études multicentriques

Dans les études multicentriques, le tirage au sort est de plus en plus souvent centralisé. Un centre, facilement joignable, sur Internet par exemple, s'occupe de la centralisation des inclusions et, après avoir vérifié le bon respect des critères d'inclusions, effectue le tirage au sort et indique à l'investigateur le traitement que doit recevoir le malade qu'il vient d'inclure. Si le tirage au sort est décentralisé, le centre qui l'a conçu, adresse à chaque centre participant à l'étude, une série d'enveloppes contenant les indications du traitement que doit recevoir chaque malade au fur et à mesure de son inclusion dans l'essai.
Des logiciels de statistiques (comme SAS®) ont des programmes de randomisation qui peuvent être déterminés en fonction du nombre de centres qui participent à l'étude, de la stratification éventuelle, du nombre de malades qui doivent être inclus, etc.
Dans ces essais multicentriques, une stratification par centre est souvent souhaitable, surtout si certains centres incluent beaucoup plus de patients que les autres.

Les risques

Rappelons qu'il y a deux risques principaux dans tout essai randomisé. Le risque de deuxième espèce doit être pris en compte dès la conception de l'étude. Il implique le calcul des effectifs de malades à inclure afin de limiter ce risque, de conclure à tort qu'un traitement n'est pas plus efficace qu'un autre alors que le nombre de malades inclus dans l'étude est insuffisant. Pendant très longtemps, l'absence de prise en compte du risque de deuxième espèce a été le grand point faible des essais randomisés. Il sera expliqué à propos des comparaisons cherchant à montrer une différence (cf. p. 92).
L'autre risque, de première espèce, se présente en fin d'étude dans la comparaison des résultats. Il consiste à conclure à tort qu'un traitement est plus efficace qu'un autre, alors que c'est essentiellement le hasard qui est intervenu dans les différences observées. Pour limiter ce risque, en pratique à moins de 5 %, il convient d'utiliser des tests d'inférence statistique (cf. p. 78).

Déviations par rapport au protocole

Il y a deux principales déviations par rapport au protocole :
- des sujets qui remplissaient les critères d'inclusion dans l'étude et qui n'ont pas été inclus ;

– des sujets qui auraient dû recevoir un traitement et qui ne l'ont pas reçu pour des raisons variées, notamment après acceptation initiale d'une allocation ou de l'autre, le refus d'un traitement.

C'est la raison pour laquelle, afin de limiter ce risque, les délais entre l'inclusion dans l'étude, le tirage au sort et la mise en œuvre de ce qu'il prévoit, doivent être aussi réduits que possible, comme il a été indiqué. En fin d'étude, il convient alors de faire une analyse des résultats en fonction du protocole alloué, puis en fonction du protocole effectué. À titre anecdotique, un essai randomisé avait été réalisé, à Boston, chez des cirrhotiques qui avaient fait au moins une hémorragie digestive par rupture de varices œsophagiennes pour comparer une dérivation porto-cave chirurgicale et la sclérose des varices dans la prévention des récidives hémorragiques [2]. Un effectif non négligeable de malades qui avaient accepté de participer à l'étude et de se faire éventuellement opérer a, secondairement, refusé l'intervention. L'analyse des résultats a montré que c'était ce sous-groupe qui avait eu les meilleurs résultats. Cela avait fait dire à l'un des hépatologues qui avaient mené cette étude que, s'il faisait une hémorragie digestive par rupture de varices œsophagiennes, il souhaiterait être inclus dans un essai randomisé, espérerait que le tirage au sort le désignerait pour être dans le groupe chirurgical et qu'il refuserait alors l'intervention.

En conclusion, les considérations précédentes montrent que chaque étape des essais randomisés demande toute une réflexion qui dépend beaucoup du rythme d'inclusion des malades dans le temps, du nombre de malades nécessaires et, dans les études multicentriques, de son organisation matérielle centralisée ou décentralisée.

Dans toute cette démarche, il faut encore avoir constamment à l'esprit les deux grandes clauses d'ambivalence et d'ignorance.

Références

1. Henk JM, Kunkler PB, Smith CW (1977) Radiotherapy and hyperbaric oxygen in head and neck cancer. Lancet 2: 101-3
2. Conn HO (1974) Therapeutic portacaval anastomosis: to shunt or not to shunt. Gastroenterology 67: 1065-73

Règles éthiques, considérations réglementaires et financement d'un essai randomisé

3

Les considérations que nous allons évoquer concernent en premier chef les essais randomisés, mais elles s'appliquent à toute évaluation diagnostique ou pronostique.

Règles éthiques

Pour un investigateur, il n'est éthique d'envisager de faire un essai randomisé pour essayer de mettre en évidence la supériorité d'un traitement sur un autre que si, l'on espère qu'un traitement est meilleur qu'un autre, mais que l'on n'en est absolument pas certain. Si on n'éprouve pas un espoir de supériorité, un essai randomisé n'a pas lieu d'être et si l'on a la conviction qu'un traitement est meilleur qu'un autre, il ne serait pas éthique non plus de mettre en œuvre un essai randomisé.
De même, il n'est éthique de faire un essai d'équivalence entre deux traitements que si l'on pense qu'il y a équivalence, mais que l'on n'en est pas certain.
Tout patient susceptible d'être inclus dans une étude prospective, doit donner son consentement éclairé ; c'est-à-dire doit être informé du but de l'étude à laquelle il accepte de participer, quels sont les avantages que l'on espère du nouveau traitement, mais aussi l'absence de certitude et les contreparties éventuelles du « nouveau » traitement comme du traitement de référence. Les patients, en général, comprennent bien l'intérêt de ces études pour la collectivité et pour eux-mêmes. La fondation d'Aide et de recherche en cancérologie a ainsi rapporté le témoignage d'une femme de 34 ans qui avait eu un cancer bilatéral des seins. Elle avait été opérée, avait eu de la chimiothérapie et de la radiothérapie. Sa tumeur n'étant pas sensible aux antihormones, on lui a proposé, pour prévenir une récidive, de participer à un essai européen. Après que les médecins se soient assurés qu'elle pouvait bien être incluse dans cette étude, elle a accepté, dit-elle, parce que si « on

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

lui proposait cet essai, c'était forcément pour améliorer ses chances de guérison ». Elle a pensé : « Au pire, que ce traitement ne serait pas efficace, mais il ne me fera pas de mal. Au mieux, il éviterait la récidive de mon cancer ».
Un autre principe éthique est que tout malade doit être prévenu qu'il est toujours libre, à tout moment, de refuser à continuer à participer à une étude.

Dispositions réglementaires[1]

Il convient de distinguer :
- les études avec bénéfice individuel direct pour le patient. Ce sont les études susceptibles (et seulement susceptibles) de lui apporter un avantage directement par rapport à des traitements antérieurs. Les essais randomisés entrent dans ce groupe ;
- et les études sans bénéfice individuel direct ; les études portant sur des sujets volontaires sains, de biodisponibilité, par exemple.
- Les premières sont le plus souvent réalisées dans des services hospitaliers, mais il en est de très utiles et intéressantes qui sont faites en pratique libérale. Les secondes, en France, doivent se dérouler dans des centres agréés par le ministère chargé de la Santé, par exemple les Centres d'investigation clinique (CIC). Le but est de protéger et de réglementer la participation les individus à de telles études contre indemnisation.

Toutes les études doivent :

1. Avoir un **promoteur** qui représente l'entité responsable sur le plan juridique et réglementaire du bon déroulement de l'étude. Lorsque l'étude a un financement propre, le promoteur est souvent le financeur. Le promoteur doit souscrire une assurance couvrant les dommages éventuels causés au patient qui participe à l'étude. Les promoteurs peuvent être des personnes physiques, mais sont le plus souvent des institutions publiques comme l'Institut national de la recherche médicale (INSERM) ou le Centre national de la recherche scientifique (CNRS) ou encore des hôpitaux, etc. D'autres sont privés, comme des firmes pharmaceutiques. Le promoteur est également responsable de la déclaration d'événements indésirables aux autorités de santé, c'est-à-dire à l'Agence française de la sécurité sanitaire pour les produits de santé (AFSSAPS).
2. Avoir un **investigateur principal** qui est le maître d'œuvre de l'étude, c'est-à-dire le coordonnateur scientifique et médical.

---

1 Directive européenne du 4 avril 2001 et loi du 9 août 2004 relative à la politique de santé publique.

3. Obtenir un **avis favorable d'un comité d'éthique**, qui est en France, un Comité pour la protection des personnes dans les recherches biomédicales (CPP). Ceux-ci sont implantés dans certains centres hospitalo-universitaires. Si un Comité refuse de donner un avis favorable, il n'est pas possible de soumettre le dossier à un autre Comité. En revanche, l'avis défavorable est habituellement assorti de recommandations dont le but est d'améliorer les conditions de l'étude. Il faut encore obtenir un avis favorable de la Commission nationale informatique et liberté (CNIL) pour les données que l'on va recueillir de façon informatique ou non.
4. Les essais prospectifs randomisés **doivent être déclarés** à l'AFSSAPS qui leur attribuent un numéro d'enregistrement. Le cas échéant, un moniteur ou un assistant de recherche clinique peut être mandaté par le promoteur et servir de lien entre lui et l'investigateur. Aux États-Unis, un site Internet[2] fournit des informations sur les essais cliniques de nouveaux traitements[3] pour faire connaître les lieux où se déroulent ces essais et leurs adresses, leur objet, et ceux qui nécessitent encore d'inclure des participants. Ces derniers ne peuvent toutefois pas s'inscrire en ligne.
La réglementation française en matière d'essai randomisé, issue de la loi sur la protection des personnes est lourde, prend du temps. Elle ne facilite pas la mise en œuvre des essais qui se font plus facilement dans certains pays étrangers. En revanche, elle a mis un terme à de petits essais conçus avec une certaine légèreté, très critiquables sur le plan méthodologique et dont on ne pouvait rien tirer ou presque des résultats. Un juste milieu reste à construire.

Financement

Les essais randomisés sont onéreux à réaliser. Leurs résultats peuvent avoir des conséquences importantes sur le plan commercial par un effet de promotion, qu'il s'agisse de médicaments, de dispositifs médicaux ou d'appareils d'exploration médicale. Ils sont de plus en plus souvent financés par des entreprises, notamment pharmaceutiques. Cela pose un réel problème de l'indépendance des investigateurs et des scientifiques qui réalisent ces essais par rapport à ceux qui les financent [1], surtout si ces derniers assurent l'enregistrement des données.
L'objectivité scientifique voudrait que le traitement des données soit fait ou bien par l'investigateur principal, ou bien par un tiers indépendant du promoteur-financeur. Au minimum, pour préserver l'indépendance

2 http://clinicaltrials:gov

3 En 2012, près de 130 000 essais dans 180 pays.

des scientifiques, ceux-ci doivent s'assurer qu'ils peuvent avoir à tout moment accès aux données, qu'ils pourront les analyser indépendamment, et qu'ils pourront préparer eux-mêmes les comptes rendus de recherche et les publier quels qu'en soient les résultats.

Enregistrement de l'essai

Tout essai randomisé destiné à être publié dans une revue internationale, doit être enregistré sur un site comme Clinicaltrials.org. Lors de cet enregistrement, le protocole sera aussi communiqué, même s'il n'est pas rendu public, pour qu'il soit possible de vérifier si l'analyse réalisée correspond bien à ce qui a été décidé dans le protocole initial. Un numéro d'enregistrement sera délivré qui devra être communiqué lors de la soumission du manuscrit décrivant l'étude.

Référence

1. Davidoff F, DeAngelis CD, Draen JM, *et al.* (2001) Sponsorship, authorship, and accountibility. N Engl J Med 345: 825-7

Comparaisons cherchant à montrer une différence (en général un bénéfice) 4

Dans toute comparaison, il y a deux risques principaux d'erreurs : le premier serait de croire qu'il y a une différence entre deux traitements alors que c'est probablement le hasard qui est intervenu dans les différences observées ; le second serait de croire qu'il n'y a pas de différence, alors que celle-ci existe. Ces comparaisons testent l'hypothèse qu'il n'y a pas de différence de bénéfice entre les deux traitements (ou examens complémentaires), dite hypothèse nulle (H0). Lorsque le test infirme cette hypothèse, on conclut qu'il y a une différence.

Le risque de première espèce

Définition

Le risque de première espèce (ou risque α) est le risque de conclure à tort, au vu des résultats, qu'un examen complémentaire, un traitement ou un facteur de pronostic est meilleur qu'un autre alors que c'est le hasard qui est responsable des différences observées. C'est le risque que prendrait un joueur à la roulette qui, sur neuf coups successifs, voyant sortir le rouge six fois et le noir trois fois, conclurait qu'il y a deux fois plus de numéros rouges que de numéros noirs.

La plupart des publications portant sur des comparaisons thérapeutiques prennent en compte ce risque de première espèce ; parfois cependant à l'aide de moyens mal adaptés, voire même inadaptés. Mais il doit être estimé dans toutes les comparaisons que ce soit des comparaisons entre deux traitements complémentaires ou entre deux facteurs de pronostic d'une maladie, etc.

Les tests d'inférence statistique sont l'outil qui permet de limiter le risque de première espèce en fixant une valeur maximale d'erreur qu'il paraît acceptable de tolérer et qui est, dans toutes les disciplines scientifiques, de 5 %.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7,

Le principe des tests statistiques : l'hypothèse nulle

Le principe général des tests statistiques, dits tests d'hypothèses, repose sur la formulation d'une hypothèse nulle, appelée hypothèse privilégiée que l'on cherche à rejeter au profit d'une hypothèse alternative. L'hypothèse nulle exprime le plus souvent l'absence de différence entre les deux éléments que l'on cherche à comparer. Rejeter cette hypothèse nulle, c'est pouvoir conclure qu'il existe une différence significative entre les deux éléments.

Pour ce faire, c'est-à-dire chercher si une différence entre les deux éléments est statistiquement significative, il convient de choisir des tests en fonction de la nature et de la distribution des variables étudiées. Le résultat de ces tests est une valeur calculée que l'on compare à des tables statistiques standardisées. Les logiciels statistiques effectuent la plupart des tests et certains sont même disponibles sur Internet (par exemple http://biostatgv.fr). Néanmoins, le principe de ces tests sera rappelé ci-après.

Comparaisons de la distribution de deux variables qualitatives : l'exemple du χ^2

Prenons l'exemple de la comparaison de la valeur diagnostique de l'angioscanner hélicoïdal (ASH) et de la résonance magnétique nucléaire (RMN) avec injection de gadolinium dans le diagnostic d'adénome hépatique. Supposons que, pour différentes raisons, la comparaison, rétrospective, ait porté sur 60 malades qui ont eu un ASH et 48 malades qui ont eu une RMN (il eut été préférable de faire un essai randomisé, ou encore mieux de faire les deux examens à chaque malade consentant) (tableau I). Les résultats observés ont montré que l'ASH a permis de faire le diagnostic chez 39 malades sur les 60 (65 %) et la RMN chez 38 malades sur les 48 (79 %). Au vu de ces pourcentages, on serait tenté de conclure que la RMN avec injection de gadolinium est un meilleur examen que l'ASH pour faire le diagnostic d'adénome hépatique. Mais cette différence peut-elle être le fait du hasard ? Autrement dit, en croyant que la RMN est supérieure à l'ASH, ne tombe-t-on pas dans le risque de première espèce ?

Pour répondre à cette question, à partir des données observées, on calcule les effectifs attendus selon l'hypothèse nulle c'est-à-dire s'il n'y avait pas de différence entre les deux examens (tableau II). Si les valeurs étaient équivalentes, la proportion de diagnostics exacts avec l'un et l'autre examen auraient été la même : 77 sur 108, soit 71 %

Tableau I – Comparaison de l'angioscanner hélicoïdal (ASH) et de la résonance magnétique nucléaire (RMN) dans le diagnostic d'adénome hépatique.

Les effectifs observés				
		ASH	**RMN**	**Total**
Diagnostic	exact	39	38	**77**
	méconnu	21	10	**31**
Total		**60**	**48**	**108**

Tableau II – Comparaison de l'angioscanner hélicoïdal (ASH) et de la résonance magnétique nucléaire (RMN) dans le diagnostic d'adénome hépatique.

Les effectifs attendus				
		ASH	**RMN**	**Total**
Diagnostic	exact	43	34	**77**
	méconnu	17	14	**31**
Total		**60**	**48**	**108**

comme le montre la troisième colonne du tableau. Rapporté aux malades qui ont eu un ASH, il y aurait eu un effectif attendu de patients chez lesquels l'ASH aurait permis de faire un diagnostic exact de 71 % de 60, soit 43 patients. Il est possible de calculer de la même façon les effectifs attendus chez les malades qui ont eu une RMN, etc. En fait, ces autres effectifs peuvent être déduits plus simplement par des soustractions à partir d'un seul effectif attendu et des totaux des lignes et des colonnes qui sont inchangés par rapport au tableau des effectifs observés.
Le test de comparaison se fait alors par le calcul du χ^2 qui mesure l'écart entre l'ensemble des effectifs observés et des effectifs attendus. Cette statistique du χ^2 est égale à la somme des carrés des différences entre chaque valeur observée et attendue, divisée par la valeur attendue (tableau III).

Tableau III – Le calcul de la valeur du χ^2.

1. Les valeurs observées				
		« Outil » diagnostic A	**« Outil » diagnostic B**	**Total**
Diagnostic	fait	*a*	*b*	*L*1
	non fait	*c*	*d*	*L*2
Total		*C*1	*C*2	*N*

4

2. Les valeurs attendues (hypothèse nulle) $a' = L1 \times C1/N$; $b' = L1 \times C2/N$, **etc.**			
	« Outil » diagnostic A	**« Outil » diagnostic B**	**Total**
Diagnostic fait	a'	b'	$L1$
non fait	c'	d'	$L2$
Total	$C1$	$C2$	N

3. Le calcul du χ²

$$\chi^2 = \frac{(a-a')^2}{(a')} + \frac{(b-b')^2}{(b')} + \frac{(c-c')^2}{(c')} + \frac{(d-d')^2}{(d')}$$

Dans notre exemple, le calcul montre que la valeur du χ^2 est égale à 2,61.
À l'aide d'une table, on calcule, à partir de cette valeur du χ^2, celle d'une valeur *p* qui estime la probabilité d'observer une telle différence par le seul effet du hasard (tableau IV). Dans notre exemple, le χ^2 étant de 2,61, on voit en lisant la première ligne de la table du χ^2 que la valeur de *p* est comprise entre 0,20 et 0,10. Comme cette valeur n'est pas inférieure à 5 % ou 0,05, on ne va pas rejeter l'hypothèse nulle, c'est-à-dire que les deux examens ont les mêmes performances. En quelque sorte, on dit que la différence observée n'est pas significative lorsqu'elle avait plus de 5 % de chances d'être obtenue par l'effet du seul hasard. Bien entendu, on aurait pu accepter une hypothèse plus laxiste, par exemple d'un *p* inférieur à 15 % au lieu de 5 %, auquel cas la valeur observée du χ^2 eut été à la limite de la signification, mais redisons-le, dans les disciplines scientifiques, la valeur maximale acceptable, unanimement admise de *p*, est de 5 %.

Tableau IV – Tableau simplifié du χ^2 pour un degré de liberté.

Valeurs du χ^2	1,07	1,64	2,71	**3,84**	5,41	6,63	10,83
Valeurs de *p*	0,30	0,20	0,10	**0,05**	0,02	0,01	0,001

Les degrés de liberté

Lorsque, comme dans notre exemple, la comparaison porte sur deux variables (diagnostic fait ou non fait) et à deux classes (ASH et RMN), connaissant les totaux des deux lignes (*L*1 et *L*2) et des deux colonnes (*C*1 et *C*2), on peut à partir d'un seul nombre du champ du tableau calculer par des soustractions les autres. On dit alors qu'il y a un degré de liberté.

Lorsque les comparaisons portent sur des variables à plus de deux classes et si les données correspondantes s'expriment dans un tableau à plus de deux colonnes (le nombre de ces colonnes étant *NC*) et à plus de deux lignes (le nombre de lignes étant *NL*), le nombre de degrés de liberté est égal à $(NC - 1) \times (NL - 1)$. Pour un risque de première espèce de 0,05, le seuil de signification des tests, χ^2 par exemple, diffère selon le nombre de degrés de liberté. La table du χ^2 (tableau V) montre que la valeur du χ^2 de 3,84 correspond à une valeur de *p* égale à 0,05 avec un degré de liberté. Cette valeur, pour deux degrés de liberté, serait de 5,99 (deuxième ligne, troisième colonne du tableau).

Tableau V – Table du χ^2.

	Valeur de *p*						
Degrés de liberté	**0,25**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**	**0,001**
1	1,323	2,706	3,841	5,024	6,635	7,879	10,83
2	2,773	4,605	5,991	7,378	9,210	10,60	13,82
3	4,108	6,251	7,815	9,348	11,34	12,84	16,27
4	5,385	7,779	9,488	11,14	13,28	14,86	18,47
5	6,626	9,236	11,07	12,83	15,09	16,75	20,52
6	7,841	10,64	12,59	14,45	16,81	18,55	22,46
7	9,037	12,02	14,07	16,01	18,48	20,28	24,32
8	10,22	13,36	15,51	17,53	20,09	21,96	26,13
9	11,39	14,68	16,92	19,02	21,67	23,59	27,88
10	12,55	15,99	18,31	20,4;8	23,21	25,19	29,59
11	13,70	17,28	19,68	21,92	24,72	26,76	31,26
12	14,85	18,55	21,03	23,34	26,22	28,30	32,91
13	15,98	19,81	22,36	24,74	27,69	29,82	34,53
14	17,12	21,06	23,68	26,12	29,14	31,32	36,12
15	18,25	22,31	25,00	27,49	30,58	32,80	37,70
16	19,37	23,54	26,30	28,85	32,00	34,27	39,25
17	20,49	.24,77	27,59	30,19	33,41	35,72	40,79
18	21,60	25,99	28,87	31,53	34,81	37,16	42,31
19	22,72	27,20	30,14	32,8.5	36,19	38,58	43,82
20	23,83	28,41	31,41	34,17	37,57	40,00	45,32

4

Valeur de *p*							
Degrés de liberté	**0,25**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**	**0,001**
21	24,93	29,62	32,67	35,48	38,93	41,40	46,80
22	26,04	30,81	33,92	36,78	40,29	42,80	48,27
23	27,14	32,01	35,17	38,08	41,64	44,18	49,73
24	28,24	33,20	36,42	39,36	.42,98	45,56	51,18
25	29,34	34,38	37,65	40,65	44,31	46,93	52,62
26	30,43	35,56	38,89	41,92	45,64	48,29	.54,05
27	31,53	36,74	40,11	43,19	46,96	49,64	55,48
28	32,62	37,92	41,34	44,46	48,28	50,99	56,89
29	33,71	39,09	42,56	45,72	49,59	52,34	58,30
30	34,80	40,26	43,77	46,98	50,89	53,67	59,70

Nous avons pris l'exemple le plus simple : celui du χ^2 qui est un test semi-paramétrique, utilisé pour comparer les distributions de deux variables qualitatives à partir d'échantillons.

Les tests paramétriques

Les tests paramétriques sont des tests qui requièrent une hypothèse sur la distribution des variables observées, c'est-à-dire que les variables étudiées suivent une distribution connue, essentiellement la loi de Laplace-Gauss. Le test paramétrique permettant la comparaison de la moyenne de deux variables quantitatives est le test *t* de Student et l'analyse de variance lorsqu'il y a plus de deux variables. L'utilisation de ces tests repose sur l'hypothèse que les variables suivent des distributions normales de même variance. Cette hypothèse est généralement raisonnable dans de nombreux cas étudiés dans les sciences de la vie. En pratique, le test est valide, plus généralement dès que les effectifs des échantillons dépassent 30 sujets.

Échantillons indépendants et appariés

Des échantillons sont appariés lorsqu'à une valeur de l'un correspond préférentiellement une valeur de l'autre. C'est, par exemple, le cas lorsque l'on a un échantillon de pression artérielle systolique et

un autre de pression artérielle diastolique mesurés chez les mêmes patients ou encore de la comparaison des taux de cholestérol sanguin avant et après traitement chez le même patient. En effet, dans ce deuxième exemple, la baisse du cholestérol après traitement est probablement dépendante de la valeur initiale de celui-ci. En épidémiologie, les études cas-témoins de sujets de même âge, de même sexe, etc. reposent habituellement sur la constitution d'échantillons appariés.
L'appariement permet de s'affranchir de la variabilité entre les individus en se focalisant sur la variabilité intra-individuelle. Lorsque des échantillons ne sont pas appariés, c'est-à-dire qu'il n'y a pas de dépendance entre les mesures, les variances sont donc plus grandes. Pour cette raison, il convient d'utiliser des tests statistiques adaptés au fait que les échantillons sont indépendants ou, au contraire, appariés.

Le choix d'un test

Le choix d'un test dépend ainsi :
– de la nature de la variable que l'on cherche à comparer, qualitative, quantitative ou censurée ;
– de sa distribution normale ou non ;
– du nombre d'échantillons que l'on cherche à comparer, selon qu'ils sont deux ou plus de deux ;
– du caractère apparié ou indépendant des échantillons.

Les tests paramétriques en fonction des variables qui sont étudiées et des effectifs			
	Variables		
	qualitatives	**quantitatives**	**censurées**
Échantillons *n* (ou groupes de patients)			
Indépendants			
$n = 2$	χ^2 ou z	t de Student ou	logrank ou Peto
$n > 2$	χ^2	Analyse de variance à un facteur	logrank ou Peto
Appariés			
$n = 2$	χ^2 de McNemar	Analyse de variance à deux facteurs	
$n > 2$	Stuart-Maxwell χ^2 de Mantel-Haenzel	Analyse de variance à deux facteurs	
De deux variables quantitatives	Coefficient de corrélation de Pearson		

4

Les tests non paramétriques

Lorsque l'on compare des variables qualitatives discontinues ou discrètes (nombre de grossesses ou d'événements indésirables d'une chimiothérapie), plus les effectifs sont faibles, moins l'hypothèse que la distribution de leurs moyennes soit normale est probable. Dans ce cas, les résultats des tests χ^2, *t*, *z*, etc. ne seront pas valides et auront trop souvent tendance à rejeter l'hypothèse nulle.
Les tests non paramétriques sont une réponse à ce problème. Ces tests ne nécessitent pas de faire d'hypothèses sur les distributions des variables, hypothèses qui sont de toute façon difficiles à vérifier. S'affranchissant des contraintes des tests paramétriques, les tests non paramétriques sont de plus en plus utilisés. Mais les tests non paramétriques sont, en général, un peu moins puissants que les tests paramétriques. Pour l'auteur d'un travail, dans des situations limites, il pourra être un peu plus facile de mettre en évidence une différence statistiquement significative avec un test paramétrique qu'avec un test non paramétrique. Il en résulte que, pour un lecteur, une valeur de *p* à la limite de la signification est plus convaincante si elle a été estimée avec un test non paramétrique qu'avec un test paramétrique car aucune hypothèse de distribution des variables n'est nécessaire pour sous-tendre ce résultat.

Les tests non paramétriques en fonction des variables qui sont étudiées et des effectifs			
	Variables		
	qualitatives	**quantitatives**	**censurées**
Échantillons *n* (ou groupes de patients)			
Indépendants			
$n = 2$	Test exact de Fischer	Wilcoxon ou Mann-Whitney	logrank
$n > 2$	Test exact de Fischer	Kruskall-Wallis	logrank
Appariés			
$n = 2$	χ^2 de McNemar	Wilcoxon signé Friedman	
$n > 2$	Cochran *Q* ou χ^2 de Mantel-Haenzel	Test de Friedman à deux facteurs	
De deux variables quantitatives	Coefficient de corrélation des rangs de Spearman et Kendall		

Les tests semi-paramétriques

Le χ^2, le χ^2 de McNemar, le logrank, sont en fait des tests semi-paramétriques. La condition d'application du test du χ^2 est d'avoir des effectifs calculés > 5. Sinon, dans le cas d'une table à deux lignes et deux colonnes, on peut utiliser le χ^2 corrigé de Yates qui est valide lorsque les effectifs théoriques sont > 3. Le z demande que np et nq soient > 5 et le test t que les distributions des variables soient normales. Le logrank qui permet de comparer des variables censurées entre elles, comme des courbes de survie, ressemble dans sa formulation mathématique au χ^2, mais sans avoir de conditions d'application : il n'y a pas d'effectifs minimums nécessaires.

Le test *t* de Student

Le test t de Student sert à comparer deux moyennes d'échantillons indépendants. Prenons l'exemple concret du résultat du dosage de la ferritine chez les nouveau-nés de deux mois selon que la mère a reçu du fer ou un placebo pendant la grossesse (exemple emprunté à [1]). Les résultats sont les suivants (tableau VI).

Tableau VI – Exemple de test de Student.

	Mère ayant reçu pendant la grossesse	
	un placebo (n = 25)	**du fer (n = 24)**
	Ferritine des nouveau-nés à 2 mois	
Moyenne (µg/L)	130	190
Variance (s^2)	4 225	9 025

De façon plus générale :

Tableau VII – Données générales pour un test de Student.

	Effectifs (échantillons)	
	Critère de jugement	
(Variable quantitative)	n_1	n_2
Moyennes	m_1	m_2
Variances	s_1^2	s_2^2

4

On commence par calculer la variance commune (tableau VIII). Son estimation est pondérée par les effectifs de chaque sous-groupe, ou plus exactement par le nombre de degrés de liberté $n - 1$.

Tableau VIII – Test de Student.

Estimation de la variance commune (S^2T)

Rappel de la formule générale :

$$s^2{}_T = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Appliquée à notre exemple, cela donne :

$$(s^2{}_T) = \frac{(25-1)4\ 225 + (24-1)9\ 025}{(25-1)+(24-1)} = 6\ 574$$

Le calcul du test t se fait ensuite de la façon suivante (tableau IX) :

Tableau IX – Calcul de la valeur du test *t*.

$$|t| = \frac{|m_1 - m_2|}{\sqrt{\frac{s^2{}_T}{n_1} + \frac{s^2{}_T}{n_2}}}$$

Appliquée à notre exemple, cela donne :

$$|t| = \frac{|130 - 190|}{\sqrt{\frac{6\ 574}{25} + \frac{6\ 574}{24}}} = \frac{60}{23{,}17} = 2{,}59$$

La valeur calculée de t (2,59) suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté, c'est-à-dire, dans notre exemple (25 – 1) + (24 – 1), soit 47. Une table (tableau X) permet alors d'estimer, à partir du *t*, la probabilité de la part du hasard dans les différences observées. Sur cette table, on voit que pour un degré de liberté de 47 (donc compris entre 40 et 60) et une valeur de t de 2,59, p est compris entre 0,01 et 0,02, ce qui permet de conclure que le traitement par le fer des femmes enceintes, pendant la grossesse, a un effet. Les résultats observés, 130 µg/L avec le placebo et 190 µg/L avec le fer montrent que cet effet est celui d'une augmentation de la valeur de la ferritine à deux mois chez le nouveau-né.

Tableau X – Table du test de Student. Formulation bilatérale.

Degrés de liberté	0,10	0,05	0,02	0,01	0,002	0,001
1	6,314	12,706	31,82	63,66	318,3	636,.6
2	2,920	4,303	6,695	9,925	22,33	31,60
3	2,353	3,182	4,541	5,841	10,21	12,92
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,893	6,869
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,408
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228.	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	l,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,610	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,552	3,850
21	1,721	2,080	2,518	2;831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792
23	1,714	2;069	2,500	2,807	3,485	3,767
24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	.2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,385	3,646
40	1,684	2,021	4,423	2,704	3,307	3,551
60	1,671	2,000	2,390	2,660	3,232	3,460
120	1,658	1,980	2,358	2,617	3,160	3,373
∞ (normale)	1,645	1,960	2,326	2,576	3,090	3,291

Ce test *t* est valable si la distribution des variables est normale. Il faut encore que, dans chacun des groupes que l'on compare, les variances soient égales. Il est possible de rendre une distribution plus normale par transformation : on remplace les valeurs par leur racine carrée ou par leur logarithme. Sinon, il convient d'utiliser un test non paramétrique (par exemple ici, le test de Wilcoxon ou de Mann-Whitney).

L'analyse de variance

L'analyse de variance (*ANalyis Of VAriance*, appelée ANOVA) sert à comparer les moyennes de variables quantitatives dans plus de deux échantillons indépendants (ANOVA à un facteur). C'est également un moyen économique d'analyser les expériences dans lesquelles deux facteurs ou plus ont été contrôlés par l'expérimentateur (ANOVA à deux facteurs par exemple). L'ANOVA permet de tester l'égalité des moyennes et non des variances (comme le libellé d'analyse de variance pourrait le laisser supposer). L'ANOVA permet, par exemple, d'interpréter la comparaison, chez des femmes qui ont une intervention gynécologique, de la qualité de vie mesurée par des scores concernant la santé générale, la santé mentale, l'état émotionnel etc. avant l'intervention, six semaines après et six mois après [2].

La variance commune (S^2_T) est la variance qui serait estimée sur la totalité de la population étudiée, tous sous-groupes que l'on cherche à comparer, confondus. La variation interéchantillons est due aux écarts entre les moyennes de chaque échantillon et la moyenne générale. Le concept de variance résiduelle (S^2_R) diffère : c'est la moyenne des variances estimées de chaque sous-groupe. Son estimation ne nécessite pas de prendre en compte les valeurs de chacune des mesures faites sur la population étudiée. On peut cependant concevoir que la variance d'un échantillon portant sur 1 000 mesures « pèse » plus lourd que celle qui ne porterait que sur 10 mesures. L'estimation de (S^2_R) est donc obtenue en pondérant les estimations dans chaque sous-groupe par les effectifs de ceux-ci ou plus exactement par le nombre de degrés de liberté (n – 1). On peut encore dire que la variation interéchantillons est liée aux facteurs que l'on souhaite comparer (des variables quantitatives), appelée pour cette raison, variation factorielle. En revanche, la variation intra-échantillon cumule les écarts de chaque valeur individuelle de la variable à leur moyenne d'échantillon. Cette dispersion provient des fluctuations aléatoires de l'échantillon. C'est la variation résiduelle (S^2_R).

L'idée d'analyser des variances pour comparer des moyennes repose sur le fait, démontrable mathématiquement, que lorsque plusieurs moyennes ne sont pas différentes entre elles (hypothèse nulle), la variance totale (S^2_T) de l'échantillon doit être égale à la variance résiduelle (S^2_R). Dans tous les autres cas, (S^2_T) diffère de (S^2_R) en lui étant supérieure.
Dans l'hypothèse nulle, et dans cette seule hypothèse, la variance totale étant égale à la valeur résiduelle, le rapport

$$F = \frac{s^2_{\mathrm{T}} - s^2_{\mathrm{R}}}{s^2_{\mathrm{R}}} \text{ est nul.}$$

Dans tous les autres cas, s^2_T est supérieur à s^2_R. Une fois calculée la valeur de *F*, comme pour le χ^2 ou le test *t*, une table de *F* donne, en fonction du nombre de degrés de liberté, les seuils au-dessus desquels cette valeur est statistiquement différente de 0, c'est-à-dire que les différences observées sont statistiquement significatives.
Supposons que l'on veuille comparer dans trois groupes de patients A, B, et C d'effectifs n_a n_b et n_c (*N* étant le nombre total d'observations), les moyennes m_a m_b et m_c, les variances s^2_a, s^2_b, et s^2_c, pour savoir si elles sont statistiquement différentes ou non. Supposons encore que, pour chaque patient, les valeurs mesurées soient pour le groupe A, x_{a1}, x_{a2}, x_{a3}, etc., pour le groupe B, x_{b1}, x_{b2}, x_{b3}, etc. et pour le groupe C x_{c1}, x_{c2}, x_{c3}, etc.
L'objectif est de calculer une valeur *F* et de voir, en se reportant à une table, si cette valeur estimée permet ou non de rejeter l'hypothèse nulle. Pour cela, il faut calculer la variance totale et la variance résiduelle (tableau XI).

Tableau XI – Calcul de la valeur du test F dans une analyse de variance (ANOVA).

La variance totale (S^2_T) est estimée par la formule :

$$S^2_{\mathrm{T}} = \frac{\Sigma_i (x_{ai} - m_a)^2 + \Sigma_i (x_{bi} - m_b)^2 + \Sigma_i (x_{ci} - m_c)^2}{N - 1}$$

La variance résiduelle est estimée par la formule :

$$S^2_{\mathrm{R}} = \frac{(n_a - 1)s^2_{a,} + (n_b - 1)s^2_{b,} + (n_c - 1)s^2_{c,}}{(n_a - 1) + (n_b - 1) + (n_c - 1)}$$

et la valeur de *F* est égale à :

$$F = \frac{s^2_{\mathrm{T}} - s^2_{\mathrm{R}}}{s^2_{\mathrm{R}}}$$

4

En pratique, le calcul de la variance résiduelle ne nécessite pas de reprendre toutes les valeurs x_{a1}, x_{a2}, $x_{a3,}$ etc. ni les variances s^2_a, s^2_b, et s^2_c, de chaque sous-groupe, pondérées par leur degré de liberté $n_a - 1$, nb – 1, nc – 1. La valeur de F est en effet égale au rapport de la variance totale, moins la variance résiduelle sur la variance résiduelle.

Le principe des tests non paramétriques

Ces tests reposent souvent sur la transformation des valeurs observées en leur rang obtenu en les classant de la plus petite à la plus grande sur les rangs. Ils s'appliquent quelle que soit la distribution de la variable dans l'échantillon. Ils s'affranchissent ainsi de la contrainte de la normalité de distribution qui est exigée pour utiliser des tests paramétriques. Prenons la comparaison d'une variable quantitative entre deux échantillons, par exemple les valeurs de la tension artérielle maximale en millimètres de mercure, chez deux groupes de malades traités par deux médicaments antihypertenseurs différents A et B, l'analyse étant réalisée à l'aide du test de Wilcoxon (tableau XII). Il convient de commencer par classer par ordre croissant les valeurs observées comme ci-après.

Tableau XII – Comparaison de deux médicaments antihypertenseurs.

Valeur de la pression artérielle (mmHg)	Anti-hypertenseur	Rang
89	A	1
96	A	2
98	B	3
101	A	4
104	B	5
106	B	6
108	B	7

Il faut ensuite faire la somme des rangs dans un groupe. Pour des saisons de simplification, il est plus aisé de choisir celui dont l'effectif est le plus petit. Dans notre exemple, très simplifié, le groupe A. Cette somme T est égale à 1 + 2 + 4. On calcule ainsi la valeur z qui est égale à 2,23 (tableau XIII).

Tableau XIII – Calcul de la valeur Z dans un test de Wilcoxon.

nA est l'effectif du groupe A (le plus petit des deux : 3 dans notre exemple).

nB est l'effectif de l'autre groupe (ici 4).

12 est une constante, quel que soit l'effectif (en effet, si *nA* et *nB* sont > 10, *Z* suit alors une loi normale).

$$z = \frac{T + 0{,}5 - nA + nB + 1)/2}{\sqrt{nA\ nB\ (nA + nB + 1)/12}} = 2{,}23$$

En se rapportant à la table du tableau du *z*, on peut rejeter l'hypothèse nulle si *z* est > 1,96, ce qui correspond à *p* < 0,05 (tableau XIV). La différence observée est alors statistiquement significative.

Tableau XIV – Table du *z*.

Valeur du Z	Probabilité (%) qu'une valeur (en valeur absolue) soit située au-delà
3,89	0,0001
3,29	0,00
2,58	0,01
2,33	0,02
2,17	0,03
2,05	0,04
1,96	0,05
1,65	0,10
1,44	0,15
1,28	0,20
1,15	0,25
1,04	0,30
0,84	0,40
0,67	0,50
0,42	0,60
0,39	0,70
0,25	0,80
0,13	0,90
0,001	0,99

Cette table signifie que la valeur d'une variable quantitative de distribution normale n'a que 5 % de chances d'être supérieure en valeur absolue à 1,96 × écart-type, c'est-à-dire a 2,5 % de chances d'être supérieure à 1,96 × écart-type et 2,5 % de chances de lui être inférieure. Lorsque les effectifs sont très petits ($nA + nB < 10$), une table spéciale doit être utilisée.

Remarques à propos du logrank [3]

Le logrank est un test non paramétrique conçu pour comparer des variables censurées. Il ne peut pas fournir d'estimation sur l'ampleur des différences observées ni donner d'intervalle de confiance ; pour ce faire, il convient d'utiliser le modèle de Cox de hasard proportionnel.

Le logrank ne s'interprète de façon simple que si la différence entre les probabilités de survie d'un groupe sont toujours de même signe, c'est-à-dire lorsque les courbes de survie ne se croisent pas.

Le risque de deuxième espèce

Ce risque (risque β) est celui de conclure à tort qu'il n'y a pas de différence entre deux examens ou deux traitements alors qu'en réalité, il y a une différence. Pour faire comprendre ce risque, prenons un exemple caricatural. On se demande si un traitement n'est pas meilleur qu'un autre sans que cela soit évident, ce qui est la condition éthique pour entreprendre un essai randomisé. Supposons que celui-ci soit parfaitement conçu à l'exception de l'estimation des effectifs de malades qu'il convient d'inclure dans l'étude. Si la comparaison ne porte que sur deux groupes de cinq malades, il est très probable que, s'il y a une différence assez faible entre les deux traitements, on ne la voie pas avec des effectifs aussi réduits, et que l'on ne rejette donc pas l'hypothèse nulle.

Beaucoup trop d'essais randomisés n'ont pas pris en compte ce risque. Il est en effet plus facile et plus rapide de mener un essai randomisé dans lequel le nombre de malades inclus est faible que s'il est important. Le résultat attendu de ces essais est qu'ils ne permettent pas de rejeter l'hypothèse nulle. Une réaction fréquente est alors de penser et de dire que les essais randomisés ne servent à rien. L'autre conséquence plus grave encore est, sur le plan médical, particulièrement préjudiciable dans le traitement d'une maladie grave, un cancer par exemple, car on ne va pas faire bénéficier les malades d'un traitement qui, pourtant, apporte une amélioration, même partielle, de

leur pronostic. Une analyse de 71 essais randomisés dont les résultats avaient été considérés comme négatifs, car non significatifs, a montré que dans 57 d'entre eux, le traitement que l'on avait évalué était, en fait, susceptible de donner des résultats 25 % meilleurs que ceux des traitements de référence, mais que les auteurs étaient passés à côté de cette différence en ne prenant pas en compte correctement le risque de deuxième espèce [4]. De plus, 34 études avaient pu méconnaître une chance d'améliorer de 50 % les résultats par rapport à ceux du traitement de référence.

Valeur acceptable du risque de deuxième espèce

Pour choisir le risque de deuxième espèce, il faut avoir préalablement défini la différence à côté de laquelle on ne voudrait pas passer si elle existait. À partir de ce choix, on fixe un seuil de risque de manquer cette différence que l'on considère comme acceptable, de la même façon que l'on se fixe le risque acceptable de première espèce de 0,05.

Il n'y a pas de valeur seuil impérative pour ce risque de deuxième espèce, notamment car ce seuil est relié à la différence que l'on souhaite ne pas manquer ou différence d'intérêt clinique. En d'autres termes, dans la comparaison d'un pourcentage observé à la valeur de 50 %, il n'y a pas de différence entre : 20 % de risque de manquer un pourcentage qui serait vraiment 69 %, 10 % de risque de manquer un pourcentage qui serait 72 %, 50 % de risque de manquer un pourcentage qui serait 64 %. En pratique, il faut donc d'abord décider quelle différence aurait un intérêt clinique si elle existait, et ensuite décider du risque que l'on est prêt à prendre de ne pas conclure à cette différence à la fin de l'essai. C'est donc la combinaison de la différence que l'on veut mettre en évidence et du risque de manquer celle-ci qui doit être considéré : 20 % de risque de manquer une petite différence peut amener à un essai plus lourd à mener, mais plus pertinent qu'un essai où l'on avait 10 % de risque de manquer une grande différence.

Souvent on prend un risque de 20 % de manquer la différence clinique d'intérêt. En effet, un plus grand risque amènerait trop souvent à une conclusion négative alors même qu'un effet existe. On peut diminuer ce risque (à 10 %, voire 5 %) si l'on souhaite réduire le risque de passer à côté d'un nouveau traitement ou test d'intérêt.

On appelle **puissance d'un test** le complément du risque β, c'est-à-dire $1 - \beta$. Un test est puissant si la probabilité de mettre en évidence une différence (rejet de l'hypothèse nulle), si différence il y a, est forte. À effectif fixé, un test est d'autant plus puissant que la différence entre les groupes est grande et à différence entre groupes fixée, un test est

d'autant plus puisant que les effectifs inclus sont grands. Se fixer un risque maximum β de 10 % veut dire que l'on se fixe une puissance maximale du test de 90 %.

La démarche du calcul de la puissance est essentiellement une démarche *a priori* lorsque l'on élabore un essai. Calculée en fin d'étude, la notion de puissance a un intérêt plus limité : si l'on a rejeté l'hypothèse nulle, on trouvera bien évidemment que le test était puissant pour la différence observée et si l'on n'a pas rejeté l'hypothèse nulle, on trouvera que la puissance était médiocre pour la différence observée.

Détermination des effectifs dans un essai randomisé

L'étape à laquelle on détermine les effectifs est souvent l'étape décisive dans la possibilité de réalisation d'un essai : de là découle l'organisation nécessaire, les ressources, le choix d'une approche multicentrique, etc. C'est aussi une étape difficile car elle va demander de faire une hypothèse, donc un pari sur ce que l'on espère gagner avec le nouveau traitement. Faire cette hypothèse est souvent l'étape pratique qui pose problème à l'investigateur. La tendance naturelle est, en effet, d'espérer que le nouveau traitement apportera plus de bénéfice qu'il n'en est, en réalité, donc de surestimer l'effet attendu.

Mais comme on l'a vu plus haut, plus cette différence supposée est grande, plus le nombre de sujets à inclure dans l'étude sera faible. Plus cette différence sera faible, plus il faut inclure de patients dans l'essai. Mais une estimation trop optimiste du bénéfice du nouveau traitement par rapport au traitement de référence entraînera un essai de petite taille et favorisera ainsi de tomber dans le risque de deuxième espèce : manquer une différence qui existait vraiment, parce qu'elle était plus petite que ce que l'on a supposé. L'excès inverse serait d'être trop pessimiste. Dans ce cas, on serait amené à inclure dans l'étude un nombre de sujets plus important que ce qui eut été nécessaire. Cela allongera d'autant la durée de l'étude et retardera les conclusions que l'on peut en tirer. Les deux envoient sur des problèmes éthiques : de trop petits essais sont peu puissants et mèneront à une absence de conclusion ; de trop grands essais constituent une perte de chances pour les patients qui auraient pu obtenir plus vite le nouveau traitement ou être orientés sur une autre thérapeutique.

Afin de ne pas s'engager à la légère dans un essai randomisé qui est toujours une entreprise lourde, il est utile de tester le nouveau traitement sur une petite série de sujets afin de se faire une première opinion de ce que l'on peut raisonnablement attendre du nouveau traitement

et fonder sur ces résultats préliminaires une hypothèse de gain qui ne soit pas trop subjective et le plus souvent trop optimiste.
Le tableau XV montre le nombre de sujets qu'il convient d'inclure dans un essai randomisé lorsque l'on compare deux pourcentages pour un risque α de 0,05 et un risque β de 0,10, c'est-à-dire une puissance du test de 90 %.

Tableau XV – Effectifs de sujets à inclure dans une comparaison de deux variables qualitatives (pour α = 0,05 et β = 0,10).

Pourcentage espéré avec le nouvel examen ou le nouveau traitement	Pourcentage connu avec l'examen ou le traitement de référence					
	5 %	10 %	20 %	30 %	40 %	50 %
10 %	578	–	263	79	59	23
15 %	184	915	1 209	158	62	33
20 %	97	263	–	389	106	48
30 %	44	79	389	–	473	121
40 %	25	39	106	473	–	515

Ce tableau appelle quatre remarques.
1. Il confirme ce qui vient d'être indiqué : plus la différence espérée est importante, plus le nombre de sujets qu'il est nécessaire d'inclure dans l'étude est faible et réciproquement. Lorsque la typhomycine a été découverte, les médecins se sont aperçus que la plupart des malades qui étaient atteints de forme grave de fièvre typhoïde et qui en mouraient, guérissaient dorénavant avec cet antibiotique. Cette observation sur un petit nombre de malades traités rendait inutile de faire un essai randomisé. Les progrès thérapeutiques sont malheureusement bien souvent moins spectaculaires au sens propre et figuré. C'est ce qui explique, justifie et nécessite à la fois la mise en œuvre d'essais randomisés et, compte tenu du nombre important de sujets qu'il est nécessaire d'inclure dans l'étude, d'être amené à faire des essais multicentriques.
2. Il existe une symétrie entre des différences (Δ) similaires. Par exemple, si le pourcentage connu est de 10 % et que le pourcentage espéré avec le nouveau « produit » est de 20 %, il faut inclure 263 sujets par groupe. Si, au contraire, le pourcentage connu était de 20 % et que l'on espère que le nouveau « produit » (par exemple une chimiothérapie moins toxique) diminuera le pourcentage de contreparties à 10 %, il faudrait inclure également 263 sujets par groupe. C'est ce

que montre encore, de façon plus générale, le calcul du nombre de sujets à inclure pour des risques donnés et en fonction des résultats connus du « produit » de référence et espéré du nouveau « produit » (tableau XVI).

3. Pour une même différence entre la valeur connue du produit de référence et du nouveau produit, plus les valeurs sont proches de 50 %, plus le nombre de sujets qu'il est nécessaire d'inclure est élevé. Dans le tableau XV, si l'on passe de 10 % à 20 % il convient d'inclure 263 sujets par groupes. Si l'on passe de 30 % à 40 %, il faut inclure 473 sujets.

Tableau XVI – Calcul du nombre de sujets à inclure dans un essai randomisé.

$$N = (1{,}96 + Z_{1-\beta})^2 \frac{2 \times \sigma^2}{\Delta^2}$$

σ est l'écart-type.

Δ est la différence espérée $|\mu_1 - \mu_2|$

Pour une puissance de 80 %, on aurait $Z_{1-\beta} = 0{,}84$

4. Plus on se fixe une puissance du test élevé (1 – β), plus il faut inclure de sujets dans l'étude, ce qui revient à dire que l'on risque moins de passer à côté d'une petite différence. En revanche, si la différence était plus importante que celle qui était pressentie, elle aurait pu être montrée plus rapidement, faisant ainsi bénéficier plus tôt les malades d'un meilleur traitement (tableau XVII).

Tableau XVII – Exemples du nombre de sujets à inclure, par groupe, pour la comparaison de deux moyennes en fonction de la puissance du test (1 – β) que l'on se fixe.

	Puissance (1 – β)			
$\lvert\mu_1 - \mu_2\rvert / \sigma$	**0,80**	**0,85**	**0,90**	**0,95**
0,10	1 571	1 797	2 102	2 600
0,30	175	201	234	290
0,50	64	73	85	105
0,70	33	38	44	54
0,90	20	23	27	33
1,10	14	16	18	22

μ_1 est la moyenne observée avec l'examen (ou traitement) de référence.
μ_1 est la moyenne espérée avec le « nouvel » examen (ou traitement) de référence.
σ est l'écart-type observé avec l'examen (ou traitement) de référence.

Le risque de troisième espèce

Ce risque (risque γ), relativement faible, mais grave par ses conséquences, est celui de conclure à tort à la supériorité d'un test diagnostique ou d'un traitement sur un autre, alors que c'est l'inverse. Pour limiter ce risque, il convient d'utiliser des tests statistiques bilatéraux (*two tailed* ou *two sided analysis*). Quand l'hypothèse nulle est rejetée, on conclut à une différence, et c'est le résultat qui montre lequel est supérieur à l'autre. Un test bilatéral implique l'inclusion d'un plus grand nombre de sujets dans l'étude.
Par exemple, une étude a comparé le traitement des carcinomes hépatocellulaires de petite taille (< 3 cm) par méthode physique percutanée et par résection chirurgicale [5]. En l'absence d'hypothèse sur celui de ces deux traitements qui était le meilleur en termes de survie, il était indispensable d'envisager dans l'interprétation des résultats un test statistique dans une formulation bilatérale.

La multiplication des tests statistiques

Dans l'interprétation des résultats d'un essai randomisé, les tests statistiques sont donc, en quelque sorte, les « garde-fous » pour faire la part du hasard et ne pas risquer de conclure, au vu d'une différence, qu'il y a réellement différence alors que c'est en grande partie le hasard qui est intervenu dans les différences observées.
Mais dans un essai, la multiplication des tests statistiques lorsqu'elle n'est pas contrôlée peut être une cause d'erreurs d'interprétation dans trois circonstances principales : c'est le cas d'analyses de résultats en cours d'essai (dites analyses intermédiaires) ; c'est encore le cas en fin d'essai, soit de l'analyse de nombreux critères de jugement secondaires, soit de l'analyse de nombreux sous-groupes de sujets inclus dans l'essai.

Les analyses intermédiaires en cours d'essai

Chez des malades qui ont eu une résection pour un cancer du poumon, si l'on veut savoir si une association de radiothérapie et de chimiothérapie, dites adjuvantes, améliore de 10 % le taux de survie à cinq ans, le faisant passer de 50 % à 60 %, pour un risque α de 0,05 et un risque β de 0,10, il faut inclure 515 malades dans chaque groupe, soit 1 030 malades en tout. Si l'on pense pouvoir inclure 150 malades par an, il faudrait un peu plus de sept ans pour mener à bien les inclusions et au moins une huitième année afin que le dernier patient inclus dans l'étude

ait au minimum un an de recul. Si l'on ajoute l'analyse des résultats, leur interprétation, le temps de rédaction d'un compte rendu de recherche, l'envoi à un périodique pour publication, les délais de réponse et de publication, il faut encore, en étant optimiste, deux années. Or, en général, on estime, compte tenu des expériences, qu'il n'est pas souhaitable, pour différentes raisons, d'entreprendre des essais pour des périodes d'inclusion des sujets dans l'étude sur plus de cinq ans.

Il est évident que, si l'on pouvait démontrer avant la fin d'un essai qu'un nouveau traitement est meilleur qu'un autre (traitement de référence ou placebo), on gagnerait un temps précieux qui éviterait de poursuivre un traitement moins bon jusqu'à la fin de l'essai, ce qui permettrait encore de faire bénéficier plus rapidement tous les malades du nouveau traitement s'il s'avérait plus efficace que le traitement de référence. La tentation est donc forte de faire des analyses intermédiaires, c'est-à-dire une analyse des résultats avant que tous les malades prévus n'aient été inclus dans l'étude. Ces analyses intermédiaires, si elles montrent une différence statistiquement significative en faveur du nouveau traitement, permettraient d'arrêter l'essai plus tôt que prévu initialement. Les analyses séquentielles (cf. p. 103) reposent sur cette idée.

Cependant, ces analyses intermédiaires **ne sont acceptables que si elles ont été prévues dans le protocole élaboré** en début d'étude, avant le commencement des inclusions. En effet, plus on augmente le nombre d'analyses intermédiaires, ne fût-ce que d'une, plus on augmente le risque que, par hasard, un test statistique montre une différence significative, c'est-à-dire, « tombe » dans le risque de première espèce. Le tableau XVIII montre, en fonction du nombre d'analyses intermédiaires, le risque global d'erreur de première espèce. De façon caricaturale, si l'on faisait un nombre infini d'analyses intermédiaires, le risque de première espèce serait de 100 %. On conclurait alors toujours à l'existence d'une différence.

Tableau XVIII – Risque global de première espèce
en fonction du nombre de tests réalisés au seuil de 5 %.

Nombre de tests réalisés	Risque global d'erreur de 1re espèce
1	0,05
2	0,08
5	0,14
10	0,19
20	0,32
∞	1,00

Pour cette raison, dans un essai randomisé, si l'on envisage de faire des analyses intermédiaires, il faut les prévoir dans le protocole initial et en tenir compte dans le calcul des effectifs afin de maintenir le risque de première espèce à 0,05. Cela implique de se fixer un premier seuil de signification du premier test que l'on réalise, plus bas que la valeur habituelle de 0,05. Dans le tableau XVIII, la valeur de 0,05 ne doit pas être la valeur du premier test, mais celle du dernier qui est prévu. Il faut donc inclure un plus grand nombre de sujets que si l'on ne faisait pas de tests intermédiaires. Il existe plusieurs règles communément utilisées : la règle de Bonferroni fixe comme seuil de signification de chaque test, non pas 0,05, mais $0{,}05/n$ où n est le nombre de tests qu'il est prévu de réaliser ; la règle de Peto fixe un seuil très faible pour les premiers tests (en général 0,0001) et réserve le gros du risque pour l'analyse finale.

Les analyses multiples en fin d'étude

Elles posent les mêmes problèmes et dans des termes analogues. Par exemple, si le but d'un essai randomisé est de comparer deux traitements pour lesquels il est prévu cinq critères de jugement indépendants, tenant compte des avantages, mais aussi des contreparties des traitements, et un seuil habituel de risque de première espèce de 0,05, le risque que l'on observe une différence significative, mais due au hasard, pour l'un des cinq critères de jugement s'élève à 14 %.

Les analyses de sous-groupes

Il en est de même pour ces analyses. Par exemple si après avoir fait un test sur l'ensemble des cas inclus dans l'essai, on fait, s'il s'agit d'un essai thérapeutique, une analyse sur le sous-groupe de malades qui n'ont pas d'envahissement ganglionnaire, une autre analyse sur le sous-groupe de malades qui ont un envahissement ganglionnaire, une autre sur ceux qui n'ont pas d'envahissement de la musculeuse s'il s'agit d'un cancer du tube digestif, etc. Il est ainsi souvent assez facile de réaliser de nombreux tests sur des sous-groupes et que l'un des tests montre une différence statistiquement significative sans tenir compte du fait qu'en multipliant le nombre des tests, on augmente d'autant le risque global d'erreur de première espèce, si l'on s'en tient au seuil « habituel » de signification de 0,05 comme le montre le tableau XVIII. Autrement dit, de la même façon que le nombre d'analyses intermédiaires augmente le risque de première espèce, il en est de même de

l'augmentation du nombre de critères de jugement. L'interprétation des différences observées concernant les critères de jugement secondaires ou des sous-groupes doit donc être d'autant plus prudente que ceux-ci sont nombreux [6], surtout si les auteurs n'ont pas pris les mêmes précautions concernant les calculs des effectifs que pour des analyses intermédiaires. Cette malfaçon est assez fréquente. Pour des auteurs qui ont fait un essai randomisé qui ne montre pas de différence significative concernant le critère de jugement principal, une manière de « sauver » leur travail est de multiplier les critères de jugement secondaires ou les analyses de sous-groupes, ce qui permet d'augmenter les chances que l'un des tests soit significatif, alors qu'il relève du risque global de première espèce.

Les réticences des comités scientifiques des périodiques médicaux et des maisons d'édition à publier des essais randomisés dont les résultats ne montrent pas de différence statistiquement significative entre les deux sujets qui ont été comparés, contribuent à favoriser ces « rattrapages » qui constituent autant de biais d'autant plus critiquables que les auteurs n'indiquent généralement pas le nombre de critères de jugement secondaires ou de sous-groupes qui ont fait l'objet de tests d'inférence statistique. Un des objectifs du dépôt des protocoles sur le site clinicaltrials.gov est d'éviter ce qui constitue une véritable malfaçon.

Références

1. Doyon F, Com-Nougué C (1983) Qu'est-ce qu'un test ? Les principaux tests statistiques. La Revue du Praticien 33: 947-54
2. Reitsma ML, Vanderkerkhof EG, Johnston SC, Hopman WM (2011) Does health-related quality of life improve in women following gynaecological surgery? J Obstet Gynaecol Can 33:1241-7
3. Bland JM (2004) The logrank test. BMJ 328: 1073 et 1412
4. Freiman JA, Chalmers TC, Smith H, Kuebler RR (2001) The importance of beta, the type II of error and sample size in the design and interpretation of the randomized controlled trial. Survey of 71 "Négative" trials. N Engl J Med 345: 825-7
5. Chen HS, Li JQ, Zheng Y, Guo RP, *et al.* (2006) A prospective randomized trial comparing percutaneous local elective therapy and partial hepatectomy for small hepatocellular carcinoma. Ann Surg 243: 21-8
6. Pocock SJ, Hughes MD, Lee RJ (1987) Statistical problems in reporting of clinical trials. A survey of medical journals. N Engl J Med 317: 426-32

Autres types d'essais randomisés

5

Essais dans lesquels les sujets sont leurs propres témoins ; essais croisés (*cross over* en anglais)

Le principe

Comme il a déjà été évoqué, dans certaines situations, il est possible que le sujet soit son propre témoin. Plusieurs éventualités existent.

C'est d'abord le cas d'examens complémentaires non invasifs, par exemple si l'on veut comparer deux examens morphologiques comme l'échographie et la résonance magnétique nucléaire et que le malade accepte d'avoir les deux examens l'un après l'autre. C'est encore le cas d'un examen biologique si le malade ne voit pas d'inconvénient à ce qu'on lui prélève un peu plus de sang pour réaliser les deux examens que l'on cherche à comparer.

Mais il y a d'autres possibilités. On peut, notamment pour des affections dermatologiques bilatérales ou diffuses, faire ou bien un traitement local d'un côté du corps et un autre traitement local sur une lésion symétrique ou bien dans une zone et dans une autre zone de dermatose.

En thérapeutique, il est encore possible de faire des essais croisés en administrant à un patient un traitement A, puis un traitement B (ou inversement), selon une séquence A-B ou B-A, tirée au sort (tableau XIX).

Tableau XIX – Schéma d'un essai croisé.

Randomisation :	Première période de l'essai	Seconde période de l'essai
Sujets 1, 3, 6	« Traitement A »	« Traitement A »
Sujets 2, 4, 5	« Traitement B »	« Traitement B »

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

C'est ce qui a été réalisé dans la comparaison de trois stratégies de traitement du diabète, traitements délivrés pendant des périodes successives de 3,5 mois à un groupe de malades afin d'étudier l'efficacité biologique et la tolérance de chaque stratégie [1].

Les avantages

Le fait que le sujet soit son propre témoin devrait, sur le plan arithmétique, permettre de diviser par deux le nombre total de sujets à inclure dans l'étude pour un seuil donné de risque de deuxième espèce, lorsqu'il y a deux groupes indépendants. Si ce nombre est *N* avec un essai classique, il devient *N/2* dans un essai croisé, mais on double le temps de participation de chaque sujet.
En réalité, sur le plan statistique, les choses sont un peu plus complexes. Si le critère de jugement est quantitatif, le nombre de sujets N' dépend aussi du coefficient de corrélation *r* qui va, rappelons-le de – 1 à + 1 entre les réponses d'un sujet aux deux traitements. N' est alors égal à :

$$N' = \frac{N}{2}(1-r)$$

Si la corrélation est positive et tend vers + 1, le nombre de sujets nécessaires diminue car la différence de réponse aux traitements sera peu variable d'un patient à l'autre et donc facilement mise en évidence. Si la corrélation est nulle, il n'y a pas d'avantage à avoir inclus les mêmes patients deux fois puisque les deux séries sont indépendantes. Le cas d'une corrélation négative des réponses au traitement chez un même individu est sans doute plus une curiosité mathématique qu'une situation réelle.
Les essais croisés apportent des informations supplémentaires à celles d'un essai classique en groupes parallèles. Ces derniers permettent de conclure qu'un traitement A est meilleur qu'un traitement B. La conclusion logique est alors de traiter les patients avec le produit A. Avec l'approche dans laquelle le malade est son propre témoin, la méthode permet de mesurer la proportion de patients ne répondant pas au traitement A, mais dont certains peuvent néanmoins bénéficier du traitement B en seconde ligne. Cette approche correspond bien à une attitude thérapeutique clinique : quand un traitement est inefficace, on en essaye un autre, même s'il peut être globalement moins efficace.

Les conditions

Dans les essais croisés, il faut, bien entendu, pouvoir évaluer l'effet du premier traitement administré avant de commencer le deuxième traitement. De plus, il ne faut pas que le traitement administré pendant la

première période interagisse avec celui de la seconde période, que ce soit une interaction positive, synergique ou négative, antagoniste. Il est possible de se prémunir de cette éventualité en séparant les deux périodes de traitement par un intervalle de temps libre appelé fenêtre thérapeutique (*wash out* en anglais). Il faut aussi, dans un essai croisé, que la maladie que l’on traite soit stable dans le temps, ce qui est paradoxal dans la mesure où le traitement que l’on évalue pendant cette période peut soulager temporairement le malade, même sans le guérir. Si le malade guérit avec le premier traitement, il devient impossible d’évaluer le second. Cela explique que de bonnes applications des essais croisés soient des affections chroniques : affections rhumatologiques, certaines maladies cutanées, maladie de Crohn, encore que dans ce dernier cas l’évolution par poussées complique l’interprétation des résultats.

L’analyse statistique

Le calcul initial du nombre de sujets nécessaires et l’analyse statistique reposent sur des tests pour séries appariées. Lorsque la variable est qualitative, les réponses aux deux traitements peuvent être toutes deux positives, toutes deux négatives ou divergentes. On ne retient pas les résultats concordants qui n’apportent pas d’information à la question posée : celle du meilleur traitement.

Les analyses séquentielles

Dans un essai randomisé, si l’on pouvait démontrer avant la fin prévue d’un essai qu’un nouveau traitement est meilleur qu’un autre (traitement de référence ou placebo), on gagnerait un temps précieux. C’est l’objectif des analyses intermédiaires dont nous avons aussi montré les contreparties. Il peut arriver que, dans un essai randomisé, les résultats sur les premiers malades inclus dans l’étude aient pu conduire à interrompre plus tôt que prévu l’essai. Ainsi, un essai randomisé a été entrepris chez des nouveau-nés qui avaient une hypertension pulmonaire persistante, comparant un groupe témoin et un groupe recevant une oxygénation à l’aide d’une membrane extracorporelle [2]. Il s’est avéré qu’il y a eu quatre décès chez dix enfants du groupe témoin qui recevaient un traitement conventionnel et aucun décès chez neuf enfants recevant une oxygénation. Cet essai ne pouvait pas être réalisé en aveugle. Bien que la différence observée chez ces dix-neuf premiers malades inclus dans l’essai ne soit pas statistiquement significative ($p = 0{,}09$), les pédiatres qui avaient entrepris cet essai ne se sont pas sentis en droit de le poursuivre et ont conclu en faveur de l’oxygénation à l’aide d’une membrane extracorporelle.

C'est un peu des observations de ce genre qui, bien que très rares, ont conduit à proposer des analyses séquentielles Leur but est de limiter l'inconvénient d'attendre la fin d'un essai randomisé classique pour conclure et donc pour recommander plus rapidement une attitude meilleure qu'une autre.

Le principe

Les analyses séquentielles consistent, comme leur nom le suggère, à inclure des malades dans l'étude par groupes de deux ou par petits groupes de nombre pairs – ce que l'on appelle une analyse séquentielle groupée – ou à faire une analyse cumulée après chaque inclusion. Chaque analyse porte sur l'ensemble des cas inclus depuis le début de l'essai.

Ce qui fait la différence avec ce que seraient des analyses intermédiaires très rapprochées et successives, est que chaque analyse est reportée sur un graphique préétabli (fig. 1). En ordonnées sont portées les différences entre les traitements qui sont comparés (Z), et en abscisse la quantité d'information cumulée proportionnelle à l'inverse de la variance (V). La pente des droites parallèles qui délimitent ces trois

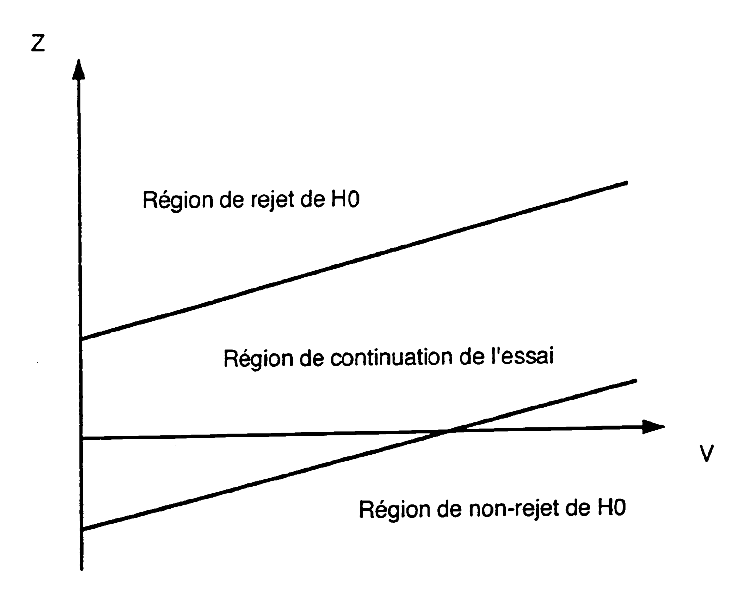


Fig. 1 – Analyse séquentielle en cas de test unilatéral. L'axe V indique la quantité d'information accumulée et l'axe Z, la différence entre les traitements comparés. Après chaque paire de malades inclus (ou groupes de paires), on reporte le résultat sur la figure jusqu'à ce que l'on sorte de la bande de continuation de l'essai.

H0 : hypothèse nulle.

zones est calculée en début d'étude selon les mêmes principes que ceux qui servent à calculer le nombre de cas qu'il est nécessaire d'inclure dans un essai randomisé classique en tenant compte des seuils choisis de risques de première et de deuxième espèce.

Après chaque analyse, un point peut ainsi être déterminé et placé sur le plan séquentiel. Lors de l'analyse suivante, à partir de l'emplacement du premier point, on place plus loin sur l'axe des abscisses un second point, etc. Le plan séquentiel comporte trois zones. Une première zone de rejet de l'hypothèse nulle correspond à la supériorité d'un traitement par rapport à un autre. Une deuxième zone est celle de l'hypothèse nulle (H0) qui doit faire poursuivre l'essai. La troisième zone est une zone dans laquelle il n'est pas possible de rejeter l'hypothèse d'absence d'égalité de traitement (en formulation unilatérale).

Dès que l'on a franchi une droite frontière, c'est-à-dire que l'on est sorti de la zone de continuation de l'essai, celui-ci est terminé. De façon générale, il est estimé que les analyses séquentielles permettent de réduire le nombre de cas à inclure dans un essai randomisé de 30 % en moyenne par rapport à une analyse unique en fin d'essai.

Néanmoins, ce type d'analyse séquentielle expose au risque de rester très longtemps dans la zone de continuation de l'essai. Pour ne pas s'exposer à cet inconvénient, il est possible de prévoir un test triangulaire, transformant les droites parallèles en un triangle (fig. 2) qui assure un nombre fini de cas à inclure dans l'essai [3].

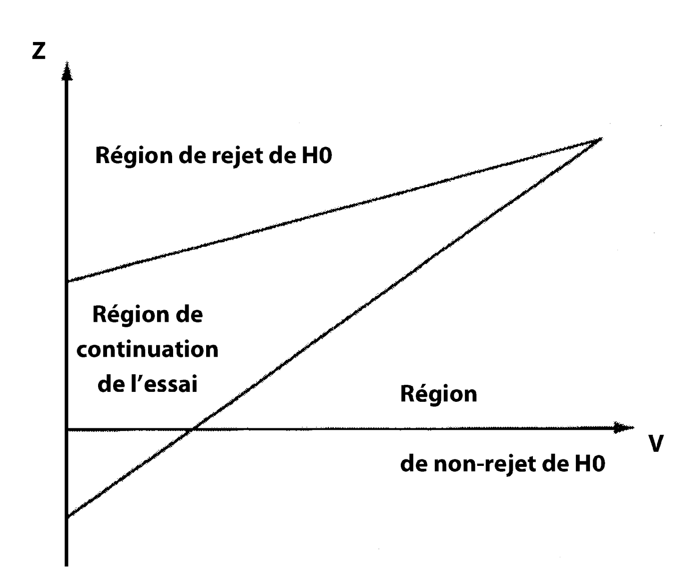


Fig. 2 – Test triangulaire en cas de test unilatéral, permettant d'éviter une poursuite indéfinie de l'essai si l'on restait dans la zone de continuation de l'essai.

H0 : hypothèse nulle.

Place des analyses séquentielles

Les analyses séquentielles sont peu utilisées, peut-être parce qu'elles se prêtent surtout à l'analyse d'un critère de jugement qui est une variable binaire. En fait, il est également possible de les adapter à des variables censurées. Cela demande une logistique plus lourde que les essais classiques, notamment parce que les analyses doivent être faites en temps réel.

Références

1. Kalergis M, Paaud D, Strychard I, *et al.* (2000) Optimizing insulin delivery: assessment of three strategies in intensive diabetes management. Diabetes Obes Metab 2: 299-305
2. O'Rourke PP, Crone RK, Vacante JP (1989) Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistant pulmonary hypertension of the newborn: a prospective randomized trial. Pediatrics 84: 957-63
3. Chastang C, Bénichou J (1992) Aspects pratiques de la planification et de l'analyse d'un essai thérapeutique randomisé selon le test triangulaire. In : Chastang C, Pons G, Régnier (eds) Méthodes nouvelles en pharmacologie clinique pédiatrique ; relation dose-effet des antibiotiques. Règles d'arrêt d'un essai clinique. Springer-Verlag, Paris, p 157-80

6 Comparaisons cherchant à montrer une équivalence

Les essais « classiques » ont pour objectif de déterminer si une innovation, que ce soit un nouvel examen radiologique ou un nouveau traitement, apporte réellement un progrès par rapport à un examen ou à un traitement antérieur, de référence. Comme nous l'avons expliqué, la méthode consiste à supposer *a priori* qu'il n'y a pas de différence entre l'ancien traitement et le nouveau : c'est l'hypothèse nulle. Si cette hypothèse est infirmée, on en déduit que le nouveau traitement est supérieur à l'ancien. Si l'hypothèse nulle est confirmée, on a tendance à déduire, de l'absence de différence significative, qu'il y a équivalence entre les deux traitements. Cette démarche est erronée : rejeter l'hypothèse d'une différence entre deux traitements ne permet pas de conclure qu'il y a équivalence entre ces deux traitements.

Or, s'assurer d'une équivalence entre deux examens ou entre deux traitements est important. Ainsi, en recherche pharmacologique de bioéquivalence, on est souvent amené à se demander si une nouvelle molécule, qui a moins de contreparties qu'une autre, entraîne le même effet biologique qu'une molécule standard antérieure. Il est encore utile de savoir si une nouvelle forme d'administration, gélule *per os*, offre la même biodisponibilité qu'un soluté injectable. En recherche clinique, on est confronté aux mêmes problèmes ; par exemple, entre une molécule princeps et un générique ou encore en oncologie entre différents modes d'administration d'une chimiothérapie dans un cancer. Toujours en oncologie, une chimiothérapie peut donner d'excellents résultats antitumoraux, mais au prix de contreparties sévères, ce qui incite à élaborer de nouveaux traitements mieux tolérés, mais dont on cherche à s'assurer qu'ils sont aussi efficaces ou du moins non inférieurs aux anciens (essais de désescalade).

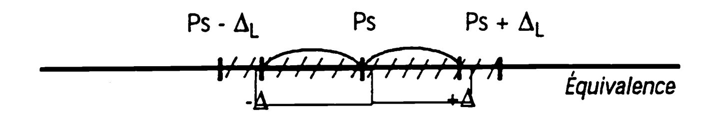
Contrairement à la démarche qui cherche à prouver la supériorité d'un traitement par rapport à un autre, dans les études d'équivalence, on part de l'hypothèse inverse : celle qu'il y a une différence d'effet entre les deux traitements. Si cette hypothèse est infirmée, on en déduit qu'il

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

y a équivalence entre ces deux traitements, c'est-à-dire que le nouveau traitement n'a pas une efficacité thérapeutique différente de celle du médicament de référence ou standard (S).

Le principe

Ainsi, l'objectif d'une étude d'équivalence est de montrer que deux traitements ne diffèrent pas en ce qui concerne le critère de jugement principal qui peut être évalué en termes de pourcentage s'il s'agit d'une variable qualitative, ou de moyenne s'il s'agit d'une variable quantitative. Le seuil maximal de différence en valeur absolue, Δ_L, des résultats des deux traitements que l'on accepte, est fixé *a priori* pour conclure qu'il y a équivalence. Il est habituellement de 10 %, mais on peut être plus exigeant ou, au contraire, moins exigeant selon les conséquences médicales et la valeur absolue du risque.
Si la différence observée $|\Delta|$ est inférieure à Δ_L, on estime qu'elle est suffisamment petite pour accepter l'hypothèse qu'il y a équivalence entre les deux traitements. En revanche, si la différence observée $|\Delta|$ est supérieure à Δ_L, on ne peut pas considérer que les deux traitements sont équivalents. C'est ce que l'on peut schématiser par la figure 1.



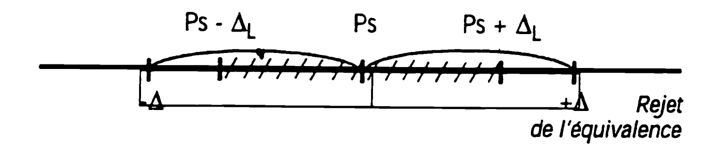
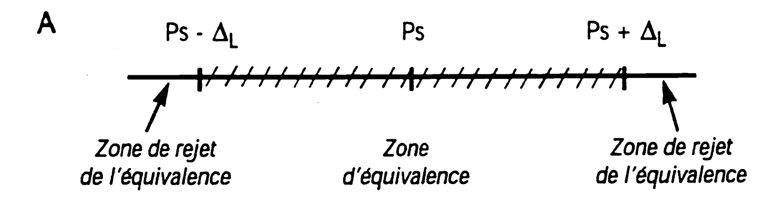


Fig. 1 – Schéma montrant le principe de détermination de l'équivalence en situation bilatérale (le nouveau médicament, s'il y a une différence avec le traitement de référence, peut être supérieur ou inférieur au médicament de référence, « produit standard » PS). Si la différence observée $|\Delta|$ est inférieure à Δ_L, on estime qu'elle est suffisamment petite pour accepter l'hypothèse qu'il y a équivalence entre les deux traitements. En revanche, si la différence observée $|\Delta|$ est supérieure à Δ_L, on ne peut pas considérer que les deux traitements sont équivalents.

En fait, deux situations peuvent se présenter. Dans le cas d'une situation bilatérale (le nouveau médicament, s'il y a une différence avec le traitement de référence, peut être supérieur ou inférieur au médicament de référence), on détermine les deux bornes de la zone d'équivalence autour de la valeur du médicament de référence (exprimée par un pourcentage ou une moyenne) (fig. 2 A). La zone d'équivalence se situe entre les deux bornes. Au-delà, il n'y a pas équivalence. Dans le cas d'une situation unilatérale, il n'y a qu'une seule borne déterminant la zone d'équivalence, située soit à gauche (fig. 2 B1), soit à droite (fig. 2 B2).



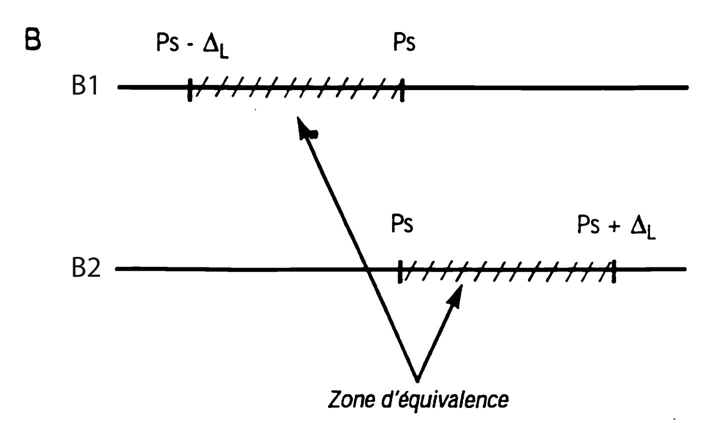


Fig. 2 – Schéma montrant le principe de détermination de l'équivalence. En A, en situation bilatérale comme dans la figure 1. En B, en situation avec une seule borne déterminant la zone d'équivalence, située soit à gauche (fig. 2 B1), soit à droite (fig. 2 B2) de la valeur (PS) du médicament de référence.

Différences entre une recherche de bénéfice (supériorité) et d'équivalence

L'hypothèse nulle (H_0)	
Il y a égalité entre A et B (A = B)	Il y a différence entre A et B (A ≠ B)
$\Delta = 0$	$\lvert\Delta\rvert \geq \Delta_L$*
L'hypothèse alternative (H_1)	
Il y a une différence entre A et B (A ≠ B)	Il y a équivalence entre A et B (A = B)
$\Delta \neq 0$	$\lvert\Delta\rvert < \Delta_L$
On teste H_0 contre H_1	
1) Si on peut rejeter H_0, on peut conclure à une différence entre A et B.	1) Si on peut rejeter H_0, on peut conclure à l'équivalence entre A et B.
2) Si on ne peut rejeter H_0, on ne peut rejeter l'hypothèse d'égalité mais on ne peut déduire qu'il y a équivalence**	2) Si on ne peut rejeter H_0, on ne peut pour autant rejeter l'hypothèse d'une différence***.

*$\lvert\Delta\rvert$ est la valeur absolue de la différence observée et Δ_L est la valeur maximale limite que l'on se fixe pour admettre l'absence de différence.
** Alors que ceux-ci peuvent être différents : problème de la puissance du test.
*** Alors que ceux-ci peuvent être équivalents : problème de la puissance du test.

Calcul du nombre de sujets nécessaires

Dans la recherche d'une supériorité d'un traitement par rapport à un autre traitement, il faut limiter le risque de conclure à tort à l'absence de différence alors que le nombre de sujets (N) inclus dans l'étude est insuffisant pour le montrer (ce qui revient à dire que la puissance du test est insuffisante). De même, dans une recherche d'équivalence (définie par $\Delta_{L)}$, il convient d'estimer le nombre de sujets nécessaires par groupe pour mettre en évidence lors d'un test de seuil α l'équivalence de deux produits qui diffèrent de moins de δ ($\delta < \Delta_L$) avec une puissance au moins égale à $(1 - \beta)$. Le calcul montre qu'il faut en général pour un essai d'équivalence un nombre de sujets comparables à ceux des essais de supériorité.

Techniques de recherche d'équivalence [1]

Pour déterminer s'il y a ou non équivalence, les tests statistiques habituels ne sont pas utilisables. Deux méthodes sont possibles.

L'une utilise l'intervalle de confiance [2]

Lorsque le critère de jugement est quantitatif, il s'agit d'une règle de décision fondée sur l'intervalle de confiance de la différence (*d*) des moyennes du produit standard (S) et du nouveau produit (N). Les limites de l'intervalle de confiance L_1 et L_2 sont fixées *a priori* et centrées sur 0 qui représente le point où la différence entre S et N est nulle. Les limites de l'intervalle de confiance sont données par le calcul :

$$d \pm \varepsilon_\alpha \text{SE}$$

où ε_α est la valeur de l'écart réduit qui est donné par une table pour un α donné, (en général 0,05, ce qui donne une valeur de ε de 1,96 et où (SE) est l'erreur standard de *d*.
Lorsque le critère de jugement est qualitatif, on raisonne sur des pourcentages au lieu de raisonner sur des moyennes. La différence (*d*) des pourcentages de bons résultats entre le produit standard (S) et le nouveau produit (N) s'écrit $d = S - N$. La déviation standard (DS) de cette différence s'obtient en calculant la racine carrée de la variance de cette différence qui n'est autre que la somme des variances de chaque pourcentage :

$$\text{Variance} = \frac{pq}{n}$$

où p est le pourcentage observé, $q = 1 - p$ et n est le nombre de sujets inclus dans l'étude.
Les limites de l'intervalle de confiance sont données, comme précédemment, par le calcul :

$$d \pm \varepsilon_\alpha \text{DS}.$$

D'autres méthodes utilisent des tests spécifiques [3]

De même que dans un essai pour chercher la supériorité d'un traitement par rapport à un autre (ou un placebo), il est nécessaire de se fixer des limites au risque que l'on accepte de prendre, notamment au risque α de première espèce de conclure à tort à l'existence d'une différence, risque que l'on cherche à infirmer.

Un exemple est tiré d'une étude nord-américaine [4]. Son but était de savoir si l'administration de soins par des infirmières n'entraînait pas de différence de résultats avec des soins administrés par des médecins. Les auteurs avaient fixé comme limite maximale de différence (Δ) acceptable, 10 %.

Trois cent quatre-vingt-douze patients ont été répartis par tirage au sort, 225 suivis par des médecins et 167 par des infirmières. Le critère de jugement a été la qualité des soins sur un certain nombre de critères aboutissant à un classement en bon ou mauvais. Le tableau I montre les résultats.

Tableau I – Résultats de l'étude. Effectifs observés (*o*) [4].

	Médecins	Infirmières	Total
Résultats :			
Bons	148	115	263
Mauvais	77	52	129
Total	$n^1 = 225$	$n^2 = 167$	N = 392

L'hypothèse testée était, ici, celle d'une différence entre résultats des médecins et ceux des infirmières. L'hypothèse alternative était celle d'une différence qui n'excède pas 10 % et que l'on choisira si le test est significatif. Le pourcentage global de soins de bonne qualité était de 263/392 = 0,67 (et celui de mauvaise qualité de 129/392 = 0,33).
Il faut commencer par calculer les effectifs théoriques des pourcentages de bons et de mauvais résultats P^1 et P^2 des infirmières et des médecins. Le calcul de P^1 et P^2 doit tenir compte du fait que les effectifs dans les deux groupes ne sont pas similaires. Il est égal à :

$$P^1 = \frac{263}{392} + \frac{167}{392} \times 0{,}10 = 0{,}71 \text{ et } P^2 = \frac{263}{392} + \frac{225}{392} \times 0{,}10 = 0{,}61$$

La différence Δ est bien de 0,10.
Les effectifs théoriques de bons résultats des médecins sont donc de 0,71 × 225 = 159,75. En reprenant les totaux de lignes et de colonne du tableau II, il est facile de calculer par complément les quatre effectifs théoriques.

Tableau II – Effectifs théoriques calculés (*c*).

	Médecins	Infirmières	Total
Résultats :			
Bons	159,75	103,75	263
Mauvais	65,25	63,75	129
Total	$n^1 = 225$	$n^2 = 167$	N = 292

Le χ^2 se calcule par la formule :

$$\chi^2 = \Sigma \frac{(o - c)^2}{c} = 6{,}48$$

La valeur de ce χ^2 traduit une différence statistiquement significative, ce qui veut dire la différence de qualité des résultats entre les médecins et les infirmières est significativement inférieure à 10 %.

Méthodes de détermination de l'équivalence

Par le calcul d'un intervalle de confiance

- *Le critère de jugement est quantitatif (moyenne)*

Par l'intervalle de confiance (IC) symétrique autour de la différence observée (d) entre les valeurs de deux traitements.

Par l'IC de la différence entre les valeurs des deux traitements, symétrique autour de 0 (exemple donné dans le texte. Westlake).

Par l'IC du rapport des valeurs de deux traitements (Mandaliaz et Mau [4]).

- *Le critère de jugement est qualitatif (pourcentage)*

Par l'IC de la différence entre les valeurs des deux traitements, symétrique autour de d.

Par l'IC de l'odds ratio.

Par des tests spécifiques

- *Le critère de jugement est quantitatif (moyenne)*

Par le test de Hauck et Anderson [5].

Par le test de Patel et Gupta [6].

- *Le critère de jugement est qualitatif (pourcentage)*

Par le test de Dunnet et Gent [7].

Conclusion

Il importe dans la formulation d'un essai de bien savoir si l'on cherche à mettre en évidence une plus grande efficacité d'un « produit » par rapport à un autre ou bien une équivalence entre des « produits ». C'est également lors de cette étape préliminaire que l'on doit définir la marge d'équivalence (Δ) au-dessous de laquelle on estime qu'une différence n'a pas suffisamment d'intérêt pour être prise en compte. Pour ce faire, on peut s'aider des études antérieures.

Ces études d'équivalence soulèvent parfois encore des problèmes d'éthique, en particulier dans les études de désescalade en cancérologie. Quelle perte de taux de succès peut-on admettre en contrepartie d'une réduction notable en toxicité ? Ou encore, quel bilan coût efficacité peut-on tenter d'établir ?

Références

1. Com-Nougue C, Rodary C (1987) Revue des procédures statistiques pour mettre en évidence l'équivalence de deux traitements. Rev Epidem Santé Publi 35: 416-30
2. Weswstlake W (1972) Use of confidence intervals in analysis of comparative biovailability trials. J Pharm Sci 61: 1340-1
3. Dunnett CW, Gent M (1977) Significance testing to establish equivalence between treatments with special reference to date in form of 2 x 2 tables. Biometrics 33: 593-602
4. Mandallaz D, Mau J (1981) Comparison of different methods for decision-making in bioequivalence assessment. Biometrics 37: 213-322
5. Hauck WW, Anderson S (1984) A new statistical procedure for testing equivalence in two-groups comparative biovailability trials. J Pharmacokin Biopharmaceut 12: 83-91
6. Patel HJ, Gupta GD (1981 March) A problem of equivalence in clinical trials. Eastern North American Region meeting. Richmond Virginia
7. Karnofski DA, Abelmann WH, Craver LF, Burchenal JH (1948) The use of nitrogen mustards in the palliative treatment of carcinoma. With particular reference to brochongenic carcinoma. Cancer 1: 634-56

Conclusions

Les malfaçons des essais randomisés

Soixante-sept essais randomisés avaient été publiés de juillet à décembre dans le *New England Journal of Medecine*, le *Lancet*, le *British Medical Journal*, ainsi que de juillet 1979 à juin 1980 dans le *Journal of the American Medical Association.* Onze critères de qualité de ces essais ont été analysés [1]. Ils étaient présents dans seulement 56 % des cas, ambigus dans 10 % et absents dans 34 %. Les critères d'inclusion n'étaient précisés que dans 19 % des cas ; la méthode de tirage au sort n'était indiquée que dans 19 % des cas ; la puissance des tests statistiques dans 12 % des cas. Il y avait des différences statistiquement significatives selon les journaux puisque les pourcentages de critères de qualité allaient de 71 % pour le *New England Journal of Medicine* à 45 % pour le *Lancet* ($p < 0{,}001$).

Un autre travail a analysé les essais randomisés publiés de juillet 1995 à juin 1998 dans les six principaux périodiques chirurgicaux en langue anglaise [2]. Les trois principales imperfections concernaient la puissance du test (68 %), la méthode de randomisation (60 %) et l'appréciation en insu ou non des critères de jugement (68 %). Il y a eu cependant une amélioration de la qualité méthodologique de ces essais au cours du temps. Elle était meilleure en 1998 que celle observée dans une étude similaire réalisée en 1981-1982.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7,

Grille d'évaluation méthodologique d'un essai randomisé [3]

Cette grille s'inspire des recommandations uniformes des comptes rendus d'essais randomisés *(Consolidated Standards of Reporting Trials CONSORT)* [4, 5].

Nous avons mis **en gras**, ce qui nous paraît à la fois particulièrement important et souvent en défaut.

1. Exposé des hypothèses qui ont motivé l'essai et son objectif.

2. Les données fondamentales :

1. Sujets inclus dans l'étude :
– critères d'inclusion et d'exclusion ;
– nombre de sujets remplissant les critères d'inclusion, mais non entrés dans l'essai et raisons ;
– description de l'échantillon.

2. Ce que l'on cherche à évaluer :
– appareil d'investigation, dispositif médical implantable, etc. (fabricant, date) ;
– ou traitement médical (posologie, mode et horaires d'administration, autres traitements admis ou non) ;
– ou traitement chirurgical (technique) ;
– ignorance en simple insu (sujet) ou en double insu (sujet et prescripteur) ;
– en cas d'événement indésirable, ce qui est prévu ?

3. Les critères de jugement :
– principal ;
– secondaires ;
– recueil par qui et comment (en insu) ?

3. Statistique

1. Calcul des effectifs
– en fonction des hypothèses médicales, des risques consentis ;
– a-t-il été prévu des analyses intermédiaires ? de sous-groupes ?

2. Randomisation
– type (permutation de nombres au hasard ?) ;
– centralisée ou non ?
– stratification ou non ?
– intervalle entre le tirage au sort et la mise en úuvre de ce que l'on cherche à évaluer.

3. Tests statistiques pertinents en fonction des variables étudiées.

4. Analyse des résultats
– déviations par rapport au protocole (inclus secondairement exclus, allocation de protocole erronée, etc.) ; jugement en intention de traiter, puis *per* protocole.
– perdus de vue
– description des groupes comparés.

5. Considérations éthiques et réglementaires
– consentement éclairé ;
– promotion et obligations légales.

6. Lors de l'élaboration du protocole
– date de début et de fin espérée des inclusions ;
– financement.

Cette grille de lecture d'un essai randomisé et de son corollaire qui est l'élaboration du protocole mérite quelques commentaires.
1) Dans la définition des données fondamentales (**2**), il ne faut jamais oublier d'envisager les interactions qui peuvent exister entre elles, ce qui amène souvent à revoir ces données à plusieurs reprises avant de les arrêter définitivement.
2) Les critères de jugement (**2, 3**) doivent être déterminés *a priori* et non au moment de l'analyse des résultats. Cela est indispensable, ne fût-ce que pour calculer les effectifs de sujets qui doivent être inclus dans l'essai.
3) Leur recueil sur des critères objectifs doit se faire, de façon préférentielle, par un observateur indépendant.
4) Le calcul de ces effectifs (**3.** 1.) est souvent difficile, négligé ou mal conduit. La prise en compte du risque de deuxième espèce est le point important qui est souvent le plus mal traité dans les essais randomisés.
5) Aucun essai n'est parfait. Faire état des déviations par rapport au protocole prévu est un signe d'honnêteté scientifique et réciproquement.

Références

1. Dersimonian R, Charrette LJ, McPeek BA, Mosteller F (1982) Reporting on methods in clinical trials. N Engl J Med 306: 1332-7
2. Schuman LP, Fischer JS, Thisted RA, Olak J (1999) Clinical trials in general surgical journals: are methods better reported? Surgery 125: 41-5
3. Charpak Y (1995) Une grille de lecture des essais thérapeutiques randomisés. Pour quoi faire ? Pour qui ? Le Concours médical 117: 2865-8
4. Altman DG (1996) Better reporting of randomised controlled trial: the CONSORT statement. Br Med J 313: 570-1
5. Liem MSL, Van der Graaf Y, Van Vroonhoven JMV (1997) CONSORT randomized trials and the scientific community. Br J Surg 84: 769-70

Partie 3

Forces d'association, études multifactorielles, mesures d'impact, causalité

Introduction

Les comparaisons entre les caractéristiques ou la réponse au traitement sont évaluées par des tests statistiques. Ces tests sont choisis en fonction de la nature de la variable étudiée, quantitative, qualitative ou censurée. Par exemple, on l'a vu (partie précédente), le test du logrank permet d'apprécier si une différence de survie entre des patients qui ont un cancer avec métastases ganglionnaires et ceux qui n'ont pas de métastases est statistiquement significative ou non.

La force d'association entre des variables relève d'un concept différent. Elle mesure l'intensité des liens qui peuvent exister entre deux ou plusieurs variables, plutôt que d'évaluer si ce lien est dû au hasard. Ces études sont dites unifactorielles (*univariate analysis* en anglais) lorsqu'elles estiment les liens entre une variable « expliquante » et une variable expliquée. Lorsqu'il existe non pas une seule, mais plusieurs variables « expliquantes », appelées covariables, les études sont dites multifactorielles *(multivariate analysis)*.

Dans ces études, les outils de mesure dépendent de la nature des variables étudiées (tableau I). Par exemple, lorsque les deux variables, à expliquer et expliquante, sont quantitatives, la force d'association est mesurée par le coefficient de corrélation ou un modèle de régression linéaire. Lorsque la variable à expliquer est qualitative et que les autres sont quantitatives ou qualitatives, cette mesure peut se faire par l'estimation des risques relatifs ou des *odds ratio*. Lorsque la variable à expliquer est censurée, on aura recours à des rapports de risques instantanés (*hazard ratio* ou HR, en anglais).

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Tableau I – Moyens d'étude des forces d'association.

	Études unifactorielles	**Études multifactorielles**
Variables « expliquées » :		
• quantitatives	régression simple	régression multiple
• qualitative	risque relatif et *odds ratio*	régression logistique
• censurée	risque relatif ; *hazard ratio*	modèle de Cox

Un problème essentiel est de ne pas confondre association et causalité. Une force d'association, statistiquement significative, n'implique pas pour autant qu'il y ait causalité entre une covariable expliquante et la variable expliquée, sauf si le dessin de l'étude permet cette explication. Un exemple simple est celui des doigts jaunis du fumeur et du cancer du poumon. Il y a une association entre les deux, cependant les doigts jaunis ne sont pas la cause du cancer du poumon. Comme nous le verrons, la causalité est difficile à mettre en évidence. Un certain nombre de critères pourront être vérifiés, mais leur absence ne peut pas faire écarter pour autant, avec certitude, un lien de causalité entre deux facteurs.

Les études unifactorielles. La régression linéaire et la corrélation

1

La corrélation

Le coefficient de corrélation de Pearson

La force d'association entre deux variables quantitatives peut être estimée par le coefficient de corrélation noté *r*. Il est égal au rapport de la covariance entre *x* et *y*, divisé par le produit de leur écart-type *s* (tableau II).

Tableau II – Le coefficient de corrélation linéaire de Pearson.

$$r = \frac{s^2_{xy}}{s_x \cdot s_y}$$

s^2_{xy} est la covariance de *X* avec *Y* et $s_{x,}$ s_y sont l'écart-type de chaque variable.

Ce coefficient peut aller de –1 à +1. S'il est supérieur en valeur absolue à 0,8, la force d'association entre les deux variables peut être considérée comme importante ; entre 0,5 et 0,8 comme modérée ; entre 0,2 et 0,5 comme faible, et très faible au dessous. Un signe positif traduit une association « positive » : la valeur de *y* croît avec celle de *x*. Une association « négative » traduit l'inverse (fig. 1). L'hypothèse nécessaire à la validité de cette mesure est que la distribution de *X* et de *Y* soient conjointement normales.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

1

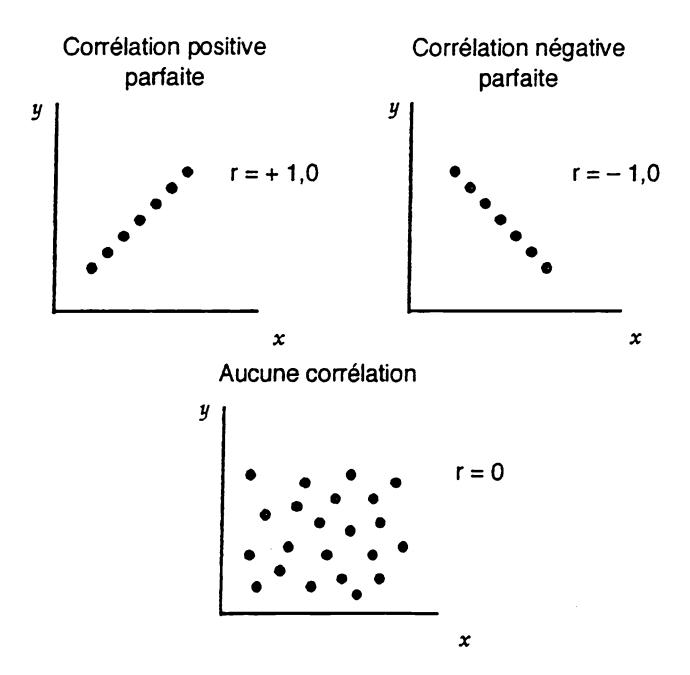


Fig. 1 – Exemple de corrélations mesurant le degré d'association entre deux variables quantitatives. Ce coefficient (r) peut aller de + 1, traduisant une corrélation positive parfaite entre deux variables, à – 1, traduisant une association totalement négative entre deux variables, en passant par 0 qui reflète l'absence de corrélation.

Tests et coefficient de corrélation

On peut tester l'hypothèse que le coefficient de corrélation r est égal à 0, c'est-à-dire l'absence d'association entre les deux variables. Si ce test est significatif, on rejette l'hypothèse nulle en concluant que les deux variables x et y ne sont pas indépendantes. Le test peut être réalisé directement à partir de la valeur de r, et dans ce cas, il est nécessaire de disposer d'une table des valeurs limites du coefficient de corrélation. Cette table est lue en fonction du nombre de degrés de liberté égale à $n-2$ ou n est le nombre de paires (x, y) analysées. Il peut aussi être réalisé en transformant la valeur de r en $t = r\sqrt{((n-2)/(1-r^2))}$, valeur qui est alors comparée au seuil de la loi de Student à $n-2$ degrés de liberté.

De la même manière qu'il existe des tests non paramétriques pour comparer des variables qualitatives entre elles, ainsi qu'entre deux variables quantitatives et qualitatives, il y a des coefficients de corrélation non paramétriques, par exemple de Kendall ou de Spierman pour les variables quantitatives. Ils permettent de s'affranchir des

contraintes d'une distribution normale, rencontrées avec le coefficient de corrélation de Pearson. Ces coefficients sont calculés à partir des rangs des observations et reposent sur le même principe que le test de Wilcoxon (cf. page 91). Leur valeur va de – 1 à 1, une valeur proche de 1 signifiant une bonne corrélation.

La régression linéaire

La régression permet d'obtenir un modèle prédictif entre deux (régression simple) ou plusieurs (régression multiple) variables. C'est donc une étape supplémentaire par rapport au coefficient de corrélation qui mesurait l'association. Lorsque les variables sont quantitatives et que la relation entre elles est linéaire, on appelle le modèle la régression linéaire. Elle est utilisée, par exemple, pour savoir si la mortalité postopératoire dans différents services de chirurgie est associée ou non au nombre d'interventions réalisées dans chaque service, dans l'année. Les données peuvent être représentées sur un graphique (fig. 2) sur lequel est porté en abscisse, pour chaque hôpital, le nombre d'interventions faites dans l'année et en ordonnée la valeur correspondante de l'autre variable, dans notre exemple, le nombre de décès postopératoires. Il en résulte un modèle statistique simple qui est la droite de régression linéaire dont l'équation s'exprime sous la forme suivante indiquée dans le tableau III.

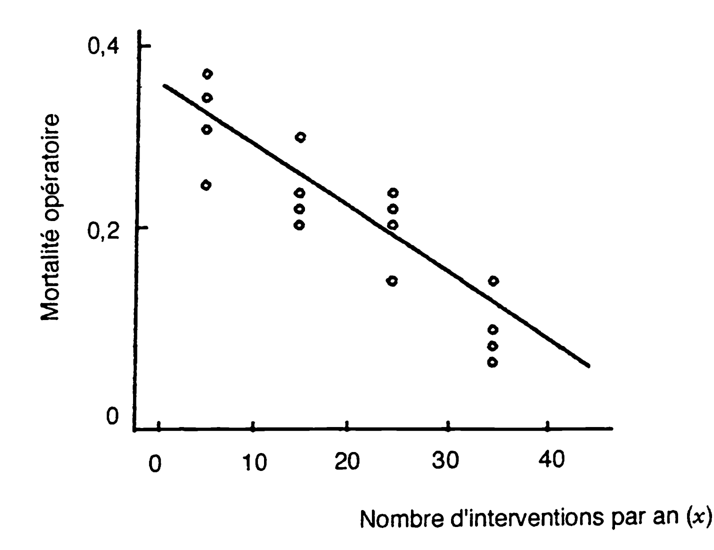


Fig. 2 – Exemple de corrélation entre deux variables quantitatives : la mortalité postopératoire et le nombre d'interventions chirurgicales réalisées dans l'année dans divers établissements hospitaliers.

Tableau III – Modèle de régression linéaire.

$$y = b + ax + e$$

Dans cette équation :

– y et x sont les deux variables quantitatives que l'on modélise, y étant la variable expliquée et x la variable expliquante ;

– b est une constante appelée ordonnée à l'origine (*intercept* en anglais) car elle représente la valeur de y lorsque x est égal à 0.

– a quantifie l'amplitude des variations de y en fonction de celles de x ; c'est la pente de la droite de régression (*slope* en anglais).

– e est un terme d'erreur, que l'on suppose de distribution normale, de moyenne nulle et de variance fixée σ^2.

Modèle prédictif

La régression permet, à partir d'une association observée, de développer un modèle prédictif, c'est-à-dire, à partir d'une valeur d'une variable x, de prédire la valeur de l'autre variable y. La droite de régression prédit la valeur moyenne de y en fonction de x. La variance de cette prédiction comprend non seulement la variance du terme d'erreur (σ^2), mais aussi un terme lié à l'incertitude d'estimation de a et b (fig. 3). Dans une telle régression, on appelle coefficient de détermination la valeur $R = r^2$. Celui-ci indique la part de la variance de Y qui est expliquée par le modèle. La valeur de R est entre 0 et 1. Plus elle est grande, plus importante est la qualité prédictive du modèle.

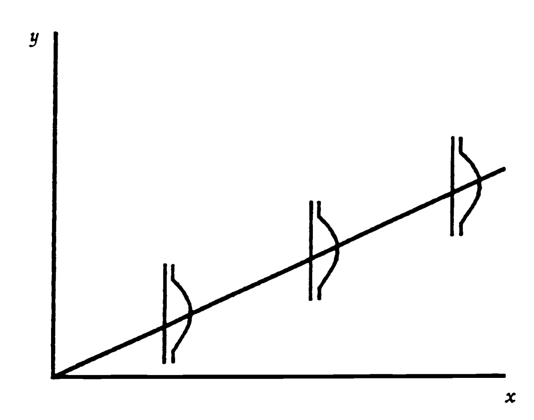


Fig. 3 – Modèle prédictif. Une droite de régression linéaire prédit la valeur moyenne de y que l'on peut assortir d'un écart-type en fonction des différentes valeurs de x.

Risques relatifs et *odds ratio*

L'étude et la connaissance des facteurs de risque sont au centre des préoccupations des épidémiologistes. Par exemple, quel est le risque de survenue d'un mésothéliome pleural chez une personne qui a été exposée à l'amiante ? Quel est le risque de survenue, l'hiver en France, d'une gastro-entérite si l'on mange des huîtres ? Cette connaissance est encore fondamentale chez un malade afin de pouvoir établir un pronostic.

Le risque absolu

Le **taux d'incidence** est une mesure descriptive en épidémiologie correspondant au nombre de nouveaux cas d'une maladie, d'une récidive ou d'un décès, recensés dans une population pendant une période de temps donnée. Autrement dit, c'est le rapport du nombre de nouveaux cas sur l'effectif de la population étudiée pendant la période donnée. **L'incidence** est une mesure du risque absolu. Par exemple, en France, sur un an, 40 nouveaux cas de cancers du côlon chez l'homme observés dans une population de 100 000 habitants correspond à une incidence annuelle de 40/100 000.

Le taux d'incidence peut aussi être estimé en personne-années. Par exemple, si 800 personnes à risque sont suivies pendant un an, et 600 autres pendant deux ans, et que 26 sont devenues séropositives au VIH pendant leur suivi, on pourra calculer un taux d'incidence de 1,3 pour 100 années-personnes. Ceci permet d'exploiter des données pour lesquelles les durées de suivi sont variées.

L'incidence permet de suivre l'évolution de la fréquence d'une affection dans le temps. Elle complète la notion de **prévalence** qui est le nombre total de cas d'une affection à un moment donné.

Le risque relatif

Le risque relatif mesure les conséquences de la présence d'un facteur de risque par rapport au risque qui existe dans une population dépourvue de ce facteur de risque. Ainsi, tout homme (ou femme) peut avoir un cancer du poumon, même s'il ne fume pas. Mais le fait de fumer augmente ce risque (et cette augmentation est dose dépendante). De même, après une intervention chirurgicale pour une maladie de Crohn, un patient qui fumait et continue à fumer a statistiquement 1,3 à 5 fois plus de risques, selon les

séries, de faire une récidive que s'il s'arrête de fumer. Inversement, la prescription d'acide 5-amino-salicylique diminue ce risque relatif [1]. Cette action bénéfique a été confirmée par des essais randomisés. Il est donc toujours nécessaire de définir une population ou une catégorie de population de référence par rapport à laquelle le risque relatif sera calculé.

Le risque relatif est mesuré par le rapport du risque absolu ou taux d'incidence chez les sujets exposés au facteur de risque sur le taux d'incidence chez les sujets qui ne sont pas exposés. Le risque relatif peut être apprécié lorsque le critère de jugement (ou événement) est qualitatif : par exemple, survenue ou non d'une gastro-entérite après consommation d'huîtres. Pour expliquer la mesure du risque relatif, nous prendrons l'exemple fictif de l'effet sur la mortalité du traitement d'un cancer par une chimiothérapie (tableau IV).

Tableau IV – Le risque relatif (RR).

Exemple fictif : effet d'une chimiothérapie sur la mortalité dans un cancer

	Patients décédés	**Patients vivants**	*Total*
Chimiothérapie	63	39	*102*
Pas de chimiothérapie	70	34	*104*
Total	*133*	*73*	*206*

Le risque de décès dans le groupe traité est de : 63/102.

Le risque de décès dans le groupe non traité est de 70/104.

Le risque relatif de décès du groupe traité par rapport au groupe non traité est de :

$$\frac{63/102}{70/104} = 0,92$$

De façon plus générale, si les données sont les suivantes :

Malades	**Non malades**	*Total*	
Exposés au risque (E)	*a*	*b*	*l*1
Non exposés (E –)	*c*	*d*	*l*2
Total	*c*1	*c*2	*N*

Le risque relatif des exposés au risque (E) par rapport au groupe non exposé (E –) est de :

$$RR = \frac{a / l1}{c / l2}$$

Un risque relatif > 1 définit un facteur de risque, et un risque relatif < 1 un facteur protecteur.

Les cotes ou *odds*

Dans les études cas-témoins, il n'est pas possible de calculer le risque absolu, et partant, les risques relatifs. Dans ces cas, on se sert des *odds* ou cotes et de leurs rapports (*odds ratio* en anglais). Le terme français de rapport de cote étant peu employé en médecine, nous utiliserons le terme anglais d'*odds ratio*. En revanche, signalons que le terme de cote est assez utilisé par les turfistes. Dans une course de chevaux, lorsque l'on dit qu'un cheval est coté 9 contre 1, c'est une cote. Elle signifie que sur 10 parieurs, 9 vont parier contre ce cheval et 1 va parier sur lui.
Les *odds ratio* sont une mesure qui approche de façon correcte le risque relatif lorsque celui-ci est faible. Même, si les *odds* et les *odds ratio* sont des notions moins intuitives que celle du risque relatif, l'*odds ratio* doit s'interpréter comme un risque relatif. Son utilisation est, avant tout, motivée par des raisons mathématiques.
Reprenons l'exemple du tableau I. L'*odds* de décès chez les malades qui ont eu de la chimiothérapie est de 63/39 soit 1,6. L'*odds* chez les malades qui n'ont pas eu de chimiothérapie est de 70/34, soit 2,1. L'*odds ratio* est le rapport de ces deux odds, soit 63 × 34/39 × 70, soit 0,8.
De façon plus générale, l'*odds ratio* est donné par la formule suivante (tableau V) :

Tableau V – L'*odds ratio.*

	Malades	**Non malades**	***Total***
Exposés au risque (E)	*a*	*b*	*l*1
Non exposés (E -)	*c*	*d*	*l*2
Total	*c*1	*c*2	*N*
L'*odds ratio* des exposés au risque (E) par rapport au groupe non exposé (E –) est de : $\frac{a/b}{c/d}=\frac{a\times d}{b\times c}$			

Remarques et interprétations des risques relatifs et des *odds ratio*

Si l'on compare ces données avec celles qui estiment le risque relatif, on se rend compte que l'*odds ratio* (0,8) se rapproche d'autant plus du risque relatif (0,9) que l'événement, ici le décès (malades du tableau IV), est rare par rapport à l'absence d'événements, ici la survie (non malades du tableau IV). *A contrario*, en termes d'efficacité d'un

traitement, l'*odds ratio* aura tendance à surestimer l'effet du traitement quand le risque de base est élevé (au-dessus de 25 % environ). Dans ces cas, les résultats représentés avec un *odds ratio* seront plus favorables au traitement que ceux fondés sur le risque relatif, et l'*odds ratio* surestimera largement le risque relatif.

Tests

Comme dans toute comparaison, il est possible, grâce à un test statistique de déterminer si le risque relatif ou l'*odds ratio* observés sont statistiquement différents de 1, c'est-à-dire de rejeter l'hypothèse d'égalité de risque ou d'*odds ratio* entre les deux éléments que l'on compare. Dans ce cas, il est possible de conclure à une association entre le facteur étudié et, dans notre exemple, la mortalité.

Intervalle de confiance

Il est aussi souhaitable d'estimer l'intervalle de confiance à 95 % autour d'un risque relatif ou d'un *odds ratio* et leur variance (s^2) (tableau VI). En fait, on ne calcule pas directement la variance, mais celle de son logarithme (Ln).

Tableau VI – Calcul de la variance d'un risque relatif ou d'un *odds ratio*.

Variance (s^2)
$s^2 (\text{Ln } \textit{odds ratio}) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$
s^2 (à partir du tableau IV) = $1/a - 1/l1 + 1/c - 1/l2$

L'intervalle de confiance à 95 % de l'*odds ratio* va de l'exponentielle de Y à l'exponentielle de Z où les valeurs de Y et Z sont :

$$Y = \text{Ln (odds ratio)} - 1{,}96 \text{ s et } Z = \text{Ln } (\textit{odds ratio}) + 1{,}96 \text{ s}$$

La formule est obtenue selon le même procédé que pour le risque relatif.

L'intervalle de confiance permet de faire le test statistique de différence à 1 du risque relatif ou de l'*odds ratio*. Dans les deux cas, si la borne inférieure de l'intervalle de confiance est supérieure à 1, le risque relatif ou l'*odds ratio* peut être dit « significativement » plus élevé que 1.

Réciproquement, si la borne supérieure de l'intervalle de confiance est inférieure à 1, on confirme statistiquement le caractère protecteur du facteur étudié.

Les analyses unifactorielles

Les analyses unifactorielles (en anglais *univariate*) consistent à estimer les liens qui peuvent exister entre une covariable (ou variable expliquante) et une variable expliquée. Par exemple, comme le montre la figure 4, dans un cancer du tube digestif, les analyses unifactorielles sur les facteurs de pronostic consistent à étudier les liens entre l'âge et la survie, puis entre l'extension pariétale du cancer et la survie, puis entre l'envahissement ganglionnaire et la survie, etc. La réalité de ces liens peut être estimée à l'aide de tests statistiques qui permettent d'apprécier si les différences observées entre l'existence ou l'absence d'une covariable comme l'envahissement ganglionnaire et la survie (ou le décès) est statistiquement significative. Les tests statistiques utilisés doivent être, bien entendu, choisis en fonction de la nature des variables étudiées (quantitatives, qualitatives ou censurées).

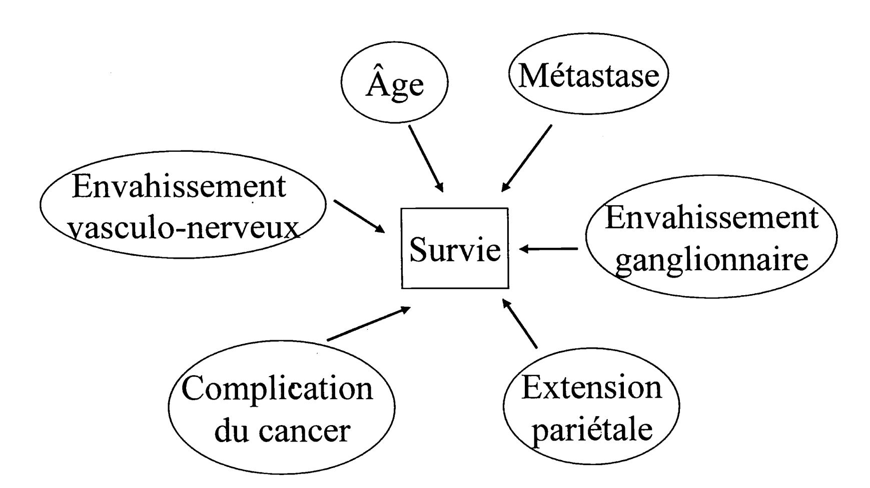


Fig. 4 – Schéma d'étude unifactorielle sur les facteurs de pronostic dans un cancer du côlon. L'étude estime pour chaque variable expliquante si le lien avec la variable expliquée, ici la survie, est statistiquement significative ou non.

La force de l'association

La force de l'association entre variables expliquantes et variable expliquée peut être estimée, selon la nature des variables par la

régression linéaire, les risques relatifs, les *odds ratio* ou les *hazard ratios.* Rappelons que si les variables expliquantes et la variable expliquée sont quantitatives, cette estimation est faite à l'aide d'un coefficient de corrélation. Si les variables sont qualitatives la force d'association est estimée à l'aide des risques relatifs ou des *odds ratio.* Si la variable à expliquer est censurée, on utilisera les *hazard ratios.*

Les limites des études unifactorielles

Les analyses unifactorielles

Les analyses unifactorielles sont très nombreuses dans la littérature. Elles concernent l'estimation soit de facteurs de pronostic dans une maladie, soit de facteurs de risque dans les études épidémiologiques de santé publique. Il convient néanmoins de bien avoir à l'esprit que la connaissance d'un facteur de risque ne permet pas pour autant de prédire une évolution vers une complication, une mortalité, ou la survenue d'une maladie. Elle permet seulement d'estimer une probabilité. Par exemple, on sait que, chez un malade qui a un cancer du côlon qui a été réséqué, l'existence d'un envahissement ganglionnaire est un facteur de mauvais pronostic, statistiquement significatif. Inversement, s'il n'existe pas d'envahissement ganglionnaire, le taux de survie à cinq ans est de 75 %. Mais chez un malade donné qui a un cancer du côlon sans métastases ganglionnaires, il n'est pas possible de savoir s'il sera parmi les 75 % de survivants ou les 25 % qui vont faire une récidive et finir par décéder de leur cancer.

Les analyses unifactorielles ont plusieurs limites.

La première, et la plus importante, est qu'elles ne tiennent pas compte des liens qui peuvent exister entre deux variables expliquantes (fig. 5). Ainsi, chez un malade qui a un cancer, l'amaigrissement, l'anorexie, l'existence d'une métastase sont liés à un mauvais pronostic. Mais ces signes sont souvent associés entre eux chez un malade : celui qui a une ou des métastases a souvent aussi un amaigrissement et une anorexie. Chacune de ces covariables n'est pas indépendante des autres. Les analyses unifactorielles peuvent apporter ainsi des informations qui sont redondantes entre elles. Elles ne permettent pas, lorsque plusieurs covariables sont statistiquement liées à la variable expliquée d'identifier celles qui le sont indépendamment des autres et qui expliquent le mieux cette variable.

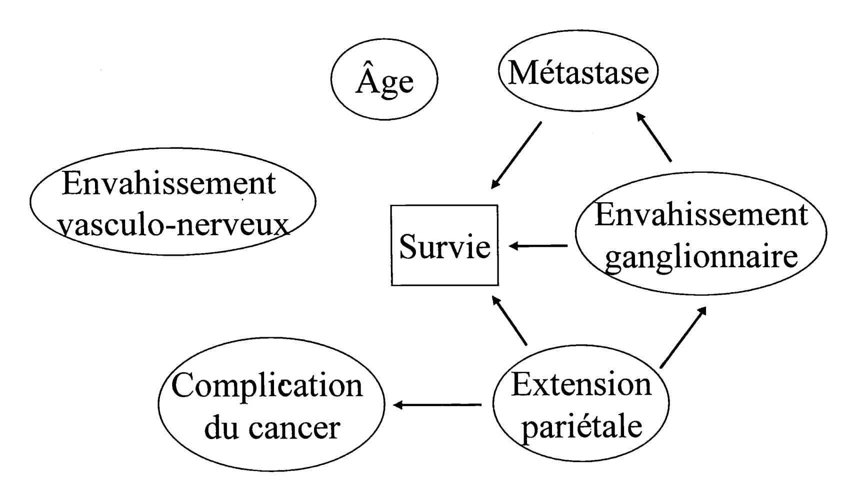


Fig. 5 – Dans les études unifactorielles, il peut y avoir des liens entre certaines variables expliquantes entre elles. Ces variables peuvent alors apporter des informations qui sont redondantes. Par exemple, plus l'extension pariétale du cancer colique est importante, plus il risque d'y avoir des métastases ganglionnaires et plus il y a de métastases ganglionnaires, plus il risque d'y avoir des métastases hépatiques.

Une autre limite des analyses unifactorielles est de ne pas permettre d'élaborer des modèles prédictifs, par exemple des scores, qui soient utiles en pratique médicale. Ainsi, une étude a montré, dans les cancers de l'estomac qui ont fait l'objet d'une résection apparemment complète, que le taux de survie à cinq ans était de 73 % en l'absence d'envahissement ganglionnaire et de 14 % lorsqu'il existait un envahissement ganglionnaire [2]. Cette même étude a encore montré des taux de survie à cinq ans de 60 % lorsque la séreuse gastrique n'était pas envahie et de 26 % si elle l'était. Si un malade a un envahissement ganglionnaire sans envahissement de la séreuse, ce qui est possible, sa probabilité de survie à cinq ans peut être estimée ainsi entre 14 % et 60 %, sans que l'on puisse être plus précis. C'est la raison pour laquelle les études unifactorielles doivent être complétées par des études multifactorielles.

Dans le même ordre d'idée, un des exemples les plus connus de ces études unifactorielles était la vieille classification de Dukes pour les cancers du rectum [3]. Cette classification reposait sur l'envahissement ou non de la musculeuse rectale, l'existence ou non de métastases ganglionnaires et l'existence ou non de métastases viscérales. Cette classification s'est avérée suffisamment bonne pour être ensuite étendue aux cancers du côlon et pour résister au temps. Elle était inspirée de deux classifications antérieures. Néanmoins, parce qu'elle reposait sur des observations unifactorielles pragmatiques, elle a posé suffisamment de problèmes pour faire l'objet de très nombreuses modifications comme le montre la figure 6 [4].

1

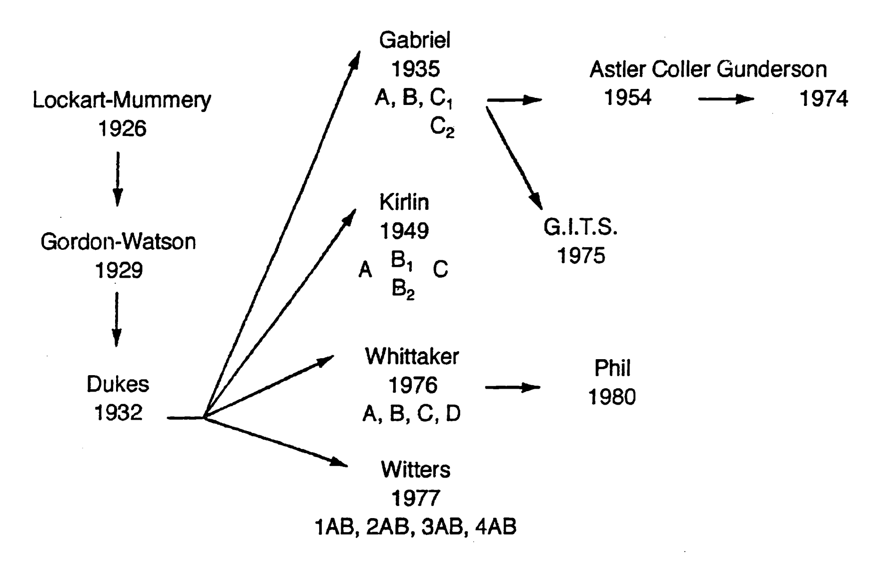


Fig. 6 – Exemple de classifications des cancers colorectaux qui, reposant sur des études unifactorielles empiriques ne se sont pas avérées satisfaisantes et ont fait l'objet de nombreuses modifications qui ont fini par mener à des confusions.

Ajustement

La réalisation d'un ajustement est l'un des aspects des études multifactorielles. Le principe de l'ajustement est de permettre de mesurer l'association entre deux variables, les autres variables étant fixées à un même niveau. Ainsi, il y a corrélation entre la pression artérielle et l'âge, et entre la pression artérielle et le poids. Le principe de l'ajustement permet de quantifier la corrélation entre la pression artérielle et l'âge comme si elles avaient été mesurées chez des individus de même poids. L'ajustement est une approche souvent utilisée lorsqu'un certain nombre de facteurs (souvent l'âge, le sexe) sont connus pour influencer le devenir du patient, mais ne constituent pas la problématique centrale d'une étude. Les techniques de l'ajustement incluent l'appariement, la stratification et la modélisation par régression.

Références

1. Borley NR, Mortensen NJ, Jewell DP (1997) Preventing postoperative recurrence of Cronh's disease. Br J Surg 84: 1493-502

2. Msika S, Chastang C, Houry S, *et al.* (1989) Lymph node involvement as the only prognostic factor in curative resected gastric carcinoma. World J Surg 12: 118-22
3. Dukes C (1932) The classification of cancer of the rectum. J Pathol Bacteriol 35: 323-32
4. Fitzgerald RH (1982) What is the Dukes' system for carcinoma of the rectum? Dis Colon Rectum 25: 774-7

Les études multifactorielles

2

Dans les sciences de la vie, s'il est intéressant d'étudier les forces d'association entre deux variables, on est beaucoup plus souvent confronté à l'étude des corrélations qu'il peut y avoir entre plusieurs covariables explicatives et une variable que l'on cherche à expliquer. C'est, par exemple, le cas des facteurs qui peuvent intervenir dans une mortalité postopératoire, facteurs liés au malade lui-même comme son âge, ses antécédents cardiovasculaires ou respiratoires, etc. ou liés à sa maladie : cancer avec ou sans métastases ganglionnaires, hépatiques ou encore du type de l'intervention réalisée.

La première étape d'une étude multifactorielle (*multivariate* en anglais) est une étude unifactorielle qui consiste, parmi toutes les variables explicatives qui ont été proposées dans l'étude, à sélectionner celles qui montrent une association statistiquement significative avec la variable expliquée. On retient habituellement pour faire cette sélection un seuil de signification qui peut être de $p = 0{,}20$ et même 0,25, c'est-à-dire supérieur au $p = 0{,}05$ habituel. Ce choix d'un seuil plus élevé a pour objectif de privilégier la puissance, c'est-à-dire de sélectionner les variables associées, même si l'association est faible, par rapport au risque de première espèce qui conduit à sélectionner des variables qui ne sont pas associées. Il peut, en effet, arriver que, combinées à d'autres variables, les associations deviennent plus fortes.

Bien entendu, une analyse multifactorielle ne peut étudier que les covariables incluses dans l'étude. L'énoncé de cette évidence a pour seul but de souligner l'importance qu'il y a de bien réfléchir au choix des covariables que l'on introduit dans le modèle. D'autre part, pour éviter d'être confronté au problème de données manquantes, il convient de faire des études prospectives.

Les analyses multifactorielles ont ainsi deux objectifs supplémentaires par rapport aux analyses unifactorielles :

1) Faire disparaître les covariables liées entre elles au profit de la seule ou des seules qui sont indépendantes. Cela équivaut à tenir compte des associations entre ces covariables. On aura cependant intérêt, lorsque plusieurs

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

covariables renseignent sur la même information à décider celle qu'il est le plus intéressant d'inclure dans le modèle. Il peut s'agir, par exemple, de celle qui est la plus simple à mesurer, la plus compréhensible, etc.

2) Permettre d'élaborer des scores prédictifs. C'est du moins le cas des méthodes qui sont dites prédictives et qui, pour cette raison, sont les plus intéressantes et les plus utiles en médecine. Dans ce cas, des variables plus simples à mesurer, et en plus petit nombre, permettront de garantir une meilleure utilisation des scores.

3) Il existe cependant des analyses multifactorielles qui reposent sur des méthodes descriptives.

S'il est possible de faire des analyses unifactorielles assez facilement ou à l'aide de logiciels simples, dès qu'il y a de nombreuses covariables, ce qui est le cas des analyses multifactorielles, l'aide de l'informatique devient indispensable.

Le principe des analyses multifactorielles

Pour faire comprendre le principe des analyses multifactorielles, nous prendrons comme exemple la connaissance du poids du nouveau-né à la naissance que l'on cherche à expliquer par l'âge et la parité de la mère (fig. 1). En effet, des études unifactorielles ont montré qu'il existait un lien entre l'âge de la mère et le poids du nouveau-né à la naissance. Plus la mère est âgée, plus le nouveau-né pèse lourd. Mais il existe aussi une liaison significative entre la parité et le poids à la naissance. Plus la parité est élevée, plus le nouveau-né pèse lourd, les

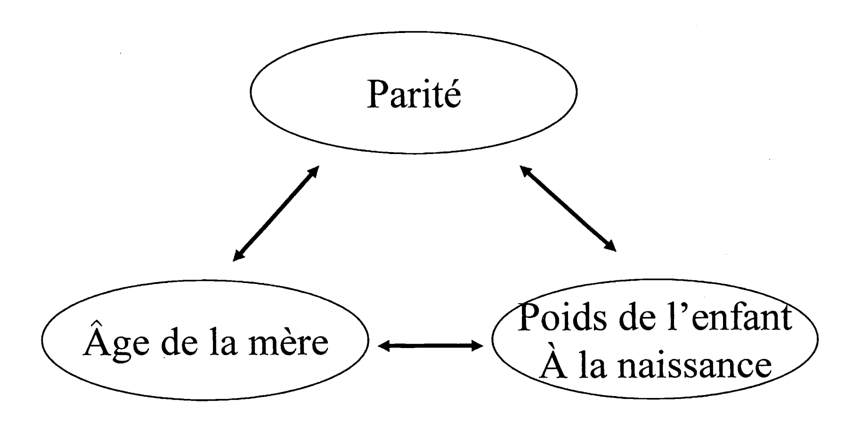


Fig. 1 – En néonatologie, il existe une corrélation entre l'âge de la mère, le poids de l'enfant à la naissance et la parité. Il existe également une corrélation entre la parité et le poids de l'enfant.

derniers nés pesant plus lourds que les premiers nés. Il existe enfin, comme cela était prévisible à partir des données précédentes, une liaison entre l'âge de la mère et la parité. On peut alors se demander si la liaison entre l'âge de la mère et le poids du nouveau-né est dépendante ou non du rang de naissance (fig. 2). Autrement dit, est-ce que la liaison entre l'âge de la mère et le poids du nouveau-né persiste ou non à parité constante ?

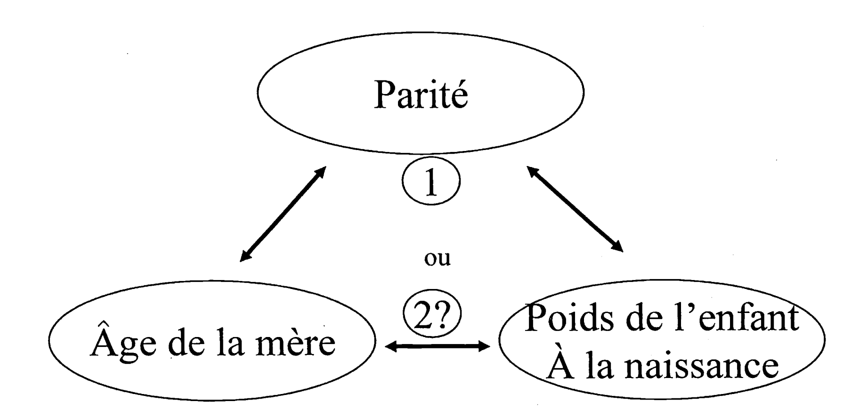


Fig. 2 – La liaison entre l'âge de la mère et le poids de la naissance de l'enfant est-elle indépendante (chemin 2) ou non (chemin 1) de la liaison avec la parité ? Les études multifactorielles permettent de répondre à cette question.

Pour cela, on recherche si le lien entre l'âge de la mère et le poids du nouveau-né est présent ou non chez les femmes primipares, puis persiste chez les secondes pares, etc. Une telle étude a montré que la liaison ne persistait pas à rang de naissance constant. La liaison entre l'âge de la mère et le poids du nouveau-né à la naissance n'était donc qu'apparente et due au fait que plus la mère est âgée, plus la parité est élevée, et plus la parité est élevée, plus les nouveau-nés pèsent lourds.

Les études multifactorielles peuvent ainsi montrer que des liaisons apparentes en analyses unifactorielles s'expliquent par un **facteur de confusion**. Dans l'exemple précédent, la parité est un facteur de confusion dans l'association poids âge. Un autre exemple bien connu est celui de l'association statiquement significative qui a été observée entre la consommation de café et le risque accru de survenue d'un infarctus du myocarde. En fait, il a aussi été observé que les fumeurs boivent plus de café que les non-fumeurs. Mais il y a encore association entre une consommation de café élevée et le tabagisme. C'est grâce à l'analyse multifactorielle que l'on a montré que le café n'augmentait pas, à lui seul, le risque d'infarctus du myocarde, mais le tabac.

Inversement, des analyses multifactorielles peuvent faire apparaître des associations statiquement significatives entre deux variables qui ne l'étaient pas en analyse unidimensionnelle, du moins de façon significative ($p \leq 0,05$).

Les analyses pas-à-pas

Les analyses multifactorielles commencent, nous l'avons indiqué, par une sélection des covariables en analyse unifactorielle. Cette sélection est fondée sur une valeur de $p \leq 0,20$, voire $\leq 0,25$. Il y a ensuite deux façons de procéder. Le pas-à-pas ascendant commence par introduire dans le modèle la covariable la plus significativement associée à la variable expliquée dans l'étude unifactorielle. C'est le pas 1. Le pas 2 consiste à introduire la covariable restante la plus associée à la variable expliquée, l'apport de la première ayant été pris en compte, et ainsi de suite. Au fur et à mesure de l'introduction d'une nouvelle covariable, c'est-à-dire à chaque pas, celle-ci contribue à l'explication de la variable expliquée, poids du nouveau-né dans notre premier exemple, infarctus du myocarde dans le second. En définitive, le modèle ne garde que les covariables associées de façon significative à la variable expliquée.
Une autre stratégie consiste, inversement, à effecteur un pas-à-pas descendant. Toutes les covariables sont initialement introduites dans le modèle (c'est le pas zéro). Les variables sont alors retirées tour à tour en partant de la moins significative. À chaque introduction d'une nouvelle covariable, si celle-ci n'est pas associée à la variable expliquée, elle est exclue du modèle. Il ne reste au dernier pas que les covariables indépendamment et significativement associées à la variable expliquée. Le pas-à-pas descendant est aujourd'hui le plus utilisé. En effet, cette stratégie privilégie des modèles plus grands, et prend en compte les facteurs de confusion avec un plus grand nombre de variables.

Importance et pertinence du choix de la population étudiée

Les études multifactorielles portant sur de très vastes échantillons qui, de ce fait, sont d'autant plus hétérogènes qu'ils sont importants, ne font parfois que confirmer ce que l'expérience simple, voire le bon sens, avaient déjà prouvé. Une très vaste étude sur les covariables associées au pronostic du cancer du poumon incluant tous les malades qui ont un cancer du poumon, va montrer qu'un mauvais pronostic est lié à l'existence de métastases, à l'étendue locorégionale de la tumeur, et à des facteurs de co-morbidité associés liés au tabagisme, ce qui n'apprend pas grand-chose.

De telles études ne sont utiles que lorsqu'elles explorent, pour la première fois, des facteurs de pronostic d'une maladie ou, en épidémiologie, des facteurs de risque d'apparition d'une maladie. Dans les autres cas, si des covariables sont connues comme étant liées à la variable que l'on cherche à expliquer, elles peuvent servir de variables d'ajustement. Ces variables sont alors un peu l'équivalent *a posteriori* de ce que nous avons vu être la stratification *a priori* dans un essai randomisé. Mais il est parfois plus intéressant de faire des études incluant uniquement des populations plus ciblées, plus homogènes, pour lesquelles la connaissance de facteurs de pronostic ou de risque a des incidences décisionnelles utiles, thérapeutiques dans le premier cas, de santé publique dans le second.

Les modèles descriptifs

L'analyse en composantes principales

L'analyse en composantes principales (ACP) traite essentiellement de variables quantitatives. Son objectif est de mettre en évidence des similarités ou des oppositions entre les covariables et à repérer celles qui sont corrélées entre elles.

L'ACP consiste à construire, à partir des variables mesurées, de nouvelles variables qui seront de variance maximale, non corrélées deux à deux et qui sont des combinaisons linéaires des variables d'origine. Ces nouvelles variables, appelées « composantes principales », peuvent servir de base à une représentation graphique des variables initiales. On peut ainsi examiner quelles sont les variables entrant dans la composition de chaque axe principal. L'interprétation des résultats se fait généralement sur les deux ou trois premiers axes principaux, sous réserve que ceux-ci expliquent la majeure partie de la variance du nuage des variables initiales. En présentant une similitude entre les variables mesurées, l'ACP est une méthode qui va permettre de réduire le nombre de variables à analyser dans un modèle multifactoriel.

L'ACP ne mesure que des liens linéaires entre variables. Avant de conclure sur l'existence ou l'absence de relations entre variables, il est donc utile d'examiner l'allure de leurs nuages de corrélation. L'ACP permet, par exemple, de résumer de nombreuses variables corrélées en une seule qui permettra l'ajustement.

L'analyse factorielle de correspondance

L'analyse factorielle de correspondance traite les variables qualitatives.

L'analyse factorielle des correspondances (AFC) ou analyse des correspondances simples est une méthode exploratoire d'analyse des tableaux de contingence. Elle vise à rassembler en un nombre réduit de dimensions la plus grande partie de l'information donnée par des tableaux de contingence, avec en tête de ligne un type de variable et en tête de colonne un ordre type de variable. L'AFC ne s'attache pas aux valeurs absolues mais aux correspondances entre les variables, c'est-à-dire aux valeurs relatives. Cette « réduction » est d'autant plus utile que le nombre de dimensions initiales est élevé. La notion de « réduction » est commune à toutes les techniques factorielles, mais l'AFC offre la particularité (contrairement aux ACP) de fournir un espace de représentation commun aux variables et aux individus.

Les modèles prédictifs

L'outil d'analyse multifactorielle dépend de la nature des variables qui sont étudiées (tableau I). Les analyses prédictives permettent, en incluant toutes les variables indépendantes et elles seules, de construire des modèles prédictifs à l'aide de score. Ces scores sont déterminés en affectant à chaque variable un coefficient plus ou moins important qui est fonction de la force d'association de chaque covariable avec la variable que l'on cherche à expliquer.

Indiquons d'emblée que ces scores demandent à être validés, soit sur des échantillons différents de ceux qui ont servi à les établir, soit par des analyses spéciales qui permettent d'apprécier leur robustesse.

Tableau I – Les différents types d'analyses multifactorielles prédictives.

Variable		**Outil**	**Expression des résultats**
« expliquante »	**« expliquée »**		
Quantitative ou qualitative ordonnée	Quantitative	Régression multiple	Coefficients de régression
Quantitative ou qualitative	Qualitative à deux classes	Régression logistique	*odds ratio*
Quantitative ou qualitative	Censurée	Modèle de Cox	Risques relatifs instantanés (*hazard ratio*)
Qualitative	Qualitative	Analyse discriminante	Valeurs prédictives

La régression linéaire multiple (*multiple linear regression* en anglais)

La régression multiple est le modèle de choix d'analyse multifactorielle lorsque les variables expliquantes et expliquées sont quantitatives. Il est encore possible de l'utiliser lorsque les covariables sont qualitatives ordonnées.
Elle permet d'analyser la valeur explicative propre de chacune des covariables étudiées. Pour ce faire, elle cherche, à l'aide de tests, à savoir si un coefficient affectant chaque covariable (coefficient de régression partielle) est différent de zéro. Si ce coefficient n'est pas différent de zéro, cela signifie que la variable correspondante n'a pas de valeur pronostique et réciproquement.
De plus, le coefficient de régression partielle mesure l'intensité de la liaison entre la covariable expliquante et la variable expliquée, à niveau constant des autres variables. Mais ce qui pouvait se calculer assez facilement, comme dans notre exemple du poids des nouveau-nés, l'âge et la parité de la mère, nécessite, dès qu'il y a de nombreuses covariables, l'aide de l'informatique.
Enfin, la régression multiple permet de trouver la combinaison linéaire de covariables permettant le mieux de décrire la variable que l'on cherche à expliquer (tableau II).

Tableau II – L'équation de régression multiple.

L'équation de régression multiple s'écrit :

$$y = a + (\beta 1 \,.\, x1) + (\beta 2 \,.\, x2) + (\beta 3 \,.\, x3) + \ldots + e.$$

dans laquelle :

– y est la variable expliquée ou dépendante ;

– a est une constante ;

– $x1$, $x2$, etc. sont les covariables expliquantes ;

– $\beta 1$, $\beta 2$ sont les coefficients de régression partielle à partir desquels il est possible de calculer la valeur de y ;

– e est l'erreur entre la valeur prédite et la valeur observée, supposée de moyenne nulle et de variance constante.

Il existe un coefficient de régression multiple R qui est calculé en tenant compte de toutes les covariables expliquantes. Il mesure la part de la variable expliquée par les covariables incluses. Plus son carré R^2 est proche de 1, mieux les covariables qui ont été incluses dans le modèle permettent de comprendre la variable expliquée. Autrement dit, le coefficient R^2 mesure le pourcentage de variabilité de la variable expliquée par les covariables étudiées, rapporté à la variabilité totale. Au point de vue pragmatique, plus R^2 est proche de 1, plus pertinent

a été le choix des covariables expliquantes et réciproquement. Dans ce dernier cas, on devrait être amené, soit à s'interroger sur le choix des covariables et à se demander si d'autres n'auraient pas été préférables, soit à prendre conscience que l'on ne sait que très imparfaitement expliquer la variable que l'on cherchait à expliquer.

Exemple

Une étude a cherché à apprécier, après différents types de résections intestinales dans la maladie de Crohn, le poids des selles et les éliminations fécales en sodium, potassium et graisses. Les covariables incluses dans le modèle ont été la longueur de chaque segment d'intestin restant : longueur restante de jéjunum, d'iléon, de côlon et de rectum [1]. Les résultats ont été exprimés par leurs coefficients de régression partielle et par des équations prédictives de l'élimination fécale, notamment en poids et en sodium (tableau III).

Tableau III – Résultats de l'analyse en régression multiple des éliminations fécales après résection intestinale en fonction de l'intestin restant et scores prédictifs [1].

	Coefficients de régression partielle			
	J	**I***	**C**	**R**
Poids fécal	– 0,47	– 0,39	– 0,33	– 0,60
P	< 0,01	< 0,01	NS	< 0,01
Sodium	– 0,24	– 0,18	– 0,36	– 0,58
P	NS	NS	< 0,01	< 0,01
	Équation prédictive de l'élimination fécale :			
Poids fécal (g) = 2,777	– 4,0 J	– 4,0 I*	– 2,5 C	– 1,2 R
Sodium (mmol) = 216	– 0,2 J	– 5,5 I*	– 0,6 C	– 100 R

J : jéjunum en centimètres ; I* : logarithme de (1 = longueur de l'iléon restant en centimètres) ; C : pourcentage de côlon restant ; R : rectum.

Les résultats d'une régression multiple peuvent être exprimés, de façon un peu préférable en indiquant pour chaque covariable, les effectifs, les coefficients de régression partielle avec leurs écarts-types et leur signification (tableau IV).

Tableau IV – Expression des résultats d'une régression linéaire multiple.

Effectifs	**Coefficients de régression partielle**	**Écart-type**	***P***
Covariable 1			
Covariable 2			
etc.			

Les variables sont classées par ordre décroissant selon l'importance de leur association avec la variable expliquée. Néanmoins, plus le nombre de covariables considérées est important, plus le risque de trouver une association « fortuite », c'est-à-dire un facteur confondant, est élevé. L'effet de ces facteurs peut cependant être corrigé.

Les coefficients de régression partielle peuvent varier de – 1 à + 1. Leurs écarts-types servent de résultat du test.

La régression logistique

La régression logistique repose sur le même principe que celui de la régression linéaire multiple. Elle est utilisable lorsque la variable expliquée est qualitative à deux classes. Les covariables étudiées peuvent être quantitatives ou qualitatives ordonnées. Pour cette raison, la régression logistique est un outil privilégié d'analyses multifactorielles.

Le tableau V montre l'équation de régression logistique.

Tableau V – L'équation de régression logistique.

L'équation de régression logistique s'écrit :

$$p(M+ \mid x1, x2, \text{etc.}) = 1 / (1 + e^{(-\beta 0 - \beta 1 \,.\, x1 - \beta 2 \,.\, x2, \text{etc.})})$$

dans laquelle $p(M+ \ldots)$ est la probabilité conditionnelle d'un événement, ici une maladie. $M+$, liée à la présence des covariables $X1$, $X2$, etc.

$\beta 0$ est une constante (*intercept* en anglais).

$\beta 1$, $\beta 2$, etc. sont les coefficients de régression partielle des variables correspondantes.

NB. Les exponentielles de $\beta 1$, $\beta 2$, etc. sont les odds ratio qui sont une approximation du risque relatif de la covariable correspondante et qui permettent d'établir des scores prédictifs.

Ici, ce ne sont pas les coefficients de régression partielle et leurs écarts-types qui servent à mesurer des associations entre les covariables et la variable expliquée, mais les *odds ratio* qui sont, rappelons-le, des approximations du risque relatif.

Une covariable dont l'*odds ratio* (ou dont le risque relatif) est égal à 1, est une covariable qui n'affecte pas le pronostic. Un facteur de bon pronostic ou un facteur protecteur se traduit par un *odds ratio* compris entre 0 et 1. Un facteur de mauvais pronostic ou un facteur de risque se traduit par un *odds ratio* supérieur à 1. Les logiciels donnent habituellement l'intervalle de confiance à 95 % des *odds ratio*. Une association est statistiquement significative lorsque l'intervalle de confiance à 95 % ne comporte pas la valeur 1. Par exemple, pour un facteur de bon pronostic, si l'*odds ratio* est de 0,4 et l'intervalle de confiance va de

0,20 à 0,70, l'association est significative. Inversement, si l'*odds ratio* est de 4,3 et l'intervalle de confiance va de 0,8 à 17,2, c'est-à-dire englobe la valeur 1, l'association n'est pas statistiquement significative. Ainsi, le fait d'indiquer l'intervalle de confiance autour d'un *odds ratio* équivaut à un test statistique.
L'expression des résultats d'une régression logistique doit exprimer pour chaque covariable les effectifs, les *odds ratio* ou les risques relatifs, l'intervalle de confiance à 95 % et le cas échant le *P* (tableau VI).

Tableau VI – Expression des résultats d'une régression logistique.

Effectifs	*Odds ratio* (ou risque relatif)	Intervalle de confiance à 95 %	*P*
Covariable 1			
Covariable 2			
etc.			

Un test d'adéquation (χ^2 *goodness of fit* d'Hosmer et Lemeshow) permet de mesurer la qualité de l'ajustement du modèle aux données. Cette mesure compare les probabilités prédites d'être un cas ou un malade, aux probabilités observées par déciles des valeurs de scores de la régression. Un bon modèle donne un test non significatif (la valeur du χ^2 étant proche de 0). Le résultat de ce test, calculé par la plupart des logiciels de statistiques est de plus en plus souvent demandé dans les publications scientifiques.

Exemple d'étude utilisant la régression logistique

Une étude a cherché à connaître, chez les malades qui avaient eu une résection-anastomose colorectale après exérèse d'un cancer du rectum, quels étaient les facteurs de risque de fistule anastomotique [2]. Seize covariables indépendantes ont été étudiées. En analyse unifactorielle, cinq d'entre elles étaient associées, de façon statistiquement significative ($p < 0{,}05$) à un risque de fistule. En analyse multifactorielle utilisant la régression logistique, seules deux covariables restaient liées au risque de fistule (tableau VII). Les résultats de cette étude ont été exprimés en termes de risque relatif assortis de leurs intervalles de confiance et de leur signification statistique.

Le modèle de Cox

Le modèle de Cox [3] repose sur le même principe général que les autres types d'analyses multifactorielles. Il est utilisé lorsque la variable expliquée est une variable censurée comme une survie, une récidive, etc.

Tableau VII – Exemple d'étude en régression logistique sur le risque de fistule anastomotique après résection du rectum pour cancer et anastomose colo-rectale [2].

	Risque relatif de fistule)	**Intervalle de confiance**	***p****
Covariable* :			
Sexe féminin	2,7	1,07 – 6,76	0,03
Anastomose < 5 cm de l'anus	6,5	2,37 – 17,87	< 0,001

Sur 16 covariables, cinq avaient été retenues après analyse unidimensionnelle, et seulement deux restaient associées au risque de fistule en analyse multidimensionnelle.
* Le *p* est un peu superfétatoire, mais montre bien que l'intervalle de confiance est l'équivalent d'un test statistique.

L'hypothèse de ce modèle est que le rapport des risques d'événements reste proportionnel au cours du temps, avec un ratio qui dépend uniquement des caractéristiques initiales des patients comparés.
La mesure de l'association fournie par le modèle est un *hazard ratio* qui estime un risque relatif instantané entre la variable expliquante et la variable expliquée. Un intervalle de confiance est calculé comme pour la régression logistique avec la même valeur statistique. Sur le plan mathématique, le modèle de Cox permet d'identifier les facteurs indépendants expliquant le risque de survenue d'un événement qui est lié au temps (tableau VIII).

Tableau VIII – Le modèle de Cox.

$h(t) = h_0\,(t) \times e^{(\beta 1\,.\,x1 + \beta 2\,.\,x2 + \beta 3\,.\,x3 \text{ etc.})}$

h(t) est le risque instantané de l'événement.

$h_0(t)$ est la fonction de risque de base, celle qui s'applique à un individu qui présenterait les niveaux de référence pour toutes les covariables du modèle.

β sont les coefficients de régression de chaque covariable *x*.

L'exponentielle de β est le risque relatif instantané (*hazard ratio*)

Exemple

Le modèle de Cox est le bon « outil » pour estimer des facteurs de pronostic d'une maladie. Une étude a ainsi été faite chez des malades qui avaient un cancer de l'œsophage, qui a été réséqué de façon apparemment complète (résection dite « à visée curative ») [4]. Sur 21 covariables analysées, en étude unidimensionnelle, neuf étaient statistiquement liées à la survie (test du logrank). Elles ont alors été incluses dans un modèle de Cox. Celui-ci a montré que seules quatre d'entre elles étaient indépendantes et associées à un mauvais pronostic (tableau IX).

Tableau IX – Facteurs de pronostic dans le cancer de l'œsophage réséqué [4].

Modèle de Cox				
Covariables	**Coefficient de régression (β)**	**Écart-type**	***P***	**Risque-relatif Instantané (HR)**
Âge < 65 ans	0,05	0,02	0,02	1,05
Classification ASA*	0,39	0,25	0,01	1,47
Infiltration pariétale	0,40	0,15	0,03	1,49
Envahissement ganglionnaire	0,38	0,19	0,01	1,46

* ASA *American society of anesthesiology*. Ce score est un score global de risque en quatre classes ordonnées qui tient compte des fonctions vitales d'un malade.

Les covariables qui dépendent du temps

Dans le modèle de Cox, les covariables doivent être appréciées au temps zéro, c'est-à-dire qui correspond à la date d'origine. Rappelons l'importance de définir une date zéro qui ait le même sens pour chacun des patients. Par exemple, dans une étude des facteurs de décès après survenue d'un infarctus du myocarde, la mesure des covariables se fera au moment de l'infarctus. Une étude de survie dans laquelle l'origine du suivi n'est pas interprétable, par exemple, la première fois que le patient est vu, quelle que soit l'étape de sa maladie, ne sera pas interprétable non plus.

Mais il peut arriver qu'une covariable, qui a une certaine valeur à la date d'origine, change ultérieurement de valeur et modifie alors le risque. Ainsi, dans les pancréatites aiguës, des facteurs de gravité peuvent apparaître seulement après quelques heures ou quelques jours d'évolution, comme la chute de l'hématocrite, une hyperleucocytose, un diabète, une élévation de la créatinémie, etc. Il est donc éminemment souhaitable que des études sur le pronostic des pancréatites aiguës prennent en compte ces données évolutives. Pour ce faire, il est possible d'utiliser un modèle adapté du modèle de Cox permettant d'inclure de telles covariables dépendant du temps.

Un exemple est la réponse à la question : la transplantation cardiaque apporte-t-elle un réel bénéfice à la survie des malades ? Comme nous l'avons vu, la meilleure réponse théorique à cette question devrait être apportée par un essai randomisé. En pratique, un tel essai, notamment pour des raisons éthiques et techniques (disponibilité de greffons), serait irréalisable. On pourrait alors se tourner vers une étude multifactorielle

en incluant, parmi les autres covariables, la transplantation. Si celle-ci était retenue par le modèle dans l'analyse multifactorielle, il serait possible de la considérer comme liée à la survie. Le problème est qu'après l'inclusion dans l'étude, des malades peuvent avoir une transplantation dans des délais qui varient beaucoup d'un patient l'autre, notamment pour des raisons de disponibilité de greffons, de biocompatibilité, etc. Apprécier les covariables seulement au moment de la transplantation elle-même risque d'introduire des biais. Par exemple, des malades à risque élevé vont mourir avant de pouvoir être transplantés. Inversement, si un malade attend deux ans sa transplantation, ce délai ne saurait être mis à l'actif de la transplantation. La meilleure façon de répondre à la question posée est donc de prendre comme date d'origine la date à laquelle l'indication de la transplantation est posée et de tenir compte ensuite de covariables dépendantes du temps comme le délai entre l'indication de la transplantation et sa réalisation.

L'analyse discriminante

L'analyse discriminante est une forme d'analyse multifactorielle dont l'objectif diffère des méthodes précédentes. Comme son nom le suggère, elle a pour but, au sein d'une population, de chercher à discriminer le mieux possible, à l'aide de covariables, deux sous-groupes A et B que différencient la survenue ou l'absence de survenue de la variable que l'on cherche à expliquer.

Si ces deux sous-groupes sont représentés par le contenu d'une ellipse, il est possible de mesurer la distance qui sépare les deux centres de ces deux ellipses (fig. 3). Plus cette distance D et son carré D^2, dénommé coefficient de Mahalanobis [5] sont importants, mieux la combinaison des covariables expliquantes discrimine les deux sous-groupes.

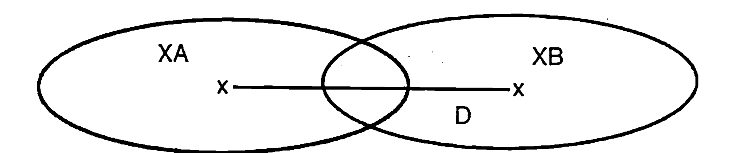


Fig. 3 – Schéma du principe de l'analyse discriminante. Plus la distance D (ou son carré D^2, appelé coefficient de Mahalanobis) est importante, mieux le modèle discrimine les deux sous-groupes, dans cet exemple, de patients décédés et survivants à un mois. Il est possible de calculer, pour une valeur donnée de D^2, la sensibilité, la spécificité et les valeurs prédictives du modèle.

Les analyses discriminantes sont, en général, effectuées pas à pas. Mais on s'aperçoit habituellement qu'après plusieurs pas, l'ajout de nouvelles covariables n'apporte proportionnellement que de moins en moins d'informations complémentaires tout en alourdissant de plus en plus le modèle. Pour ce faire, il est possible d'estimer la proportion de sujets bien classés au fur et à mesure de l'introduction des covariables expliquantes. Il est encore possible, comme nous le verrons, de calculer la sensibilité, la spécificité et mieux encore, les valeurs prédictives du modèle au fur et à mesure de l'introduction de nouvelles covariables.
L'expression des résultats d'une analyse discriminante est indiquée dans le tableau X.

Tableau X – Expression des résultats d'une analyse discriminante.

Covariables	Coefficient de Mahalanobis	*P*	% de sujets bien classés	Valeurs prédictives
1				
2				
etc.				

Plus le coefficient de Mahalanobis est élevé, plus la covariable correspondante ou l'association de covariables discrimine les deux sous-groupes que l'on cherche à identifier.

Exemple

Une étude a cherché, chez des malades cirrhotiques qui avaient fait une hémorragie digestive, liée à une hypertension portale, les facteurs de mortalité (ou de survie) à un mois [6]. Pour chaque covariable, il a d'abord été fait une analyse discriminante unidimensionnelle en évaluant le coefficient de Mahanalobis et sa signification statistique (tableau XI). Ces covariables ont été classées par ordre décroissant de ce coefficient, c'est-à-dire en commençant par les covariables qui discriminaient le plus les survivants, des malades décédés. Les auteurs ont calculé pour chaque covariable le pourcentage de malades bien classés. Comme on le voit, s'il y a une certaine cohérence entre un coefficient de Mahalanobis élevé, la signification du test et le pourcentage de bien classés, cette cohérence n'est pas absolue. Par exemple, l'ascite qui était associée au coefficient de Mahalanobis le plus fort ne vient qu'au troisième rang du classement des malades. Il aurait encore été possible d'estimer la valeur prédictive de décès ou de survie de chaque covariable.

Tableau XI – Exemple d'analyse discriminante sur la mortalité un mois après une hémorragie digestive chez les patients cirrhotiques.

Analyse discriminante unidimensionnelle			
Covariables	**Coefficient de Mahalanobis**	***P***	**% de sujets bien classés**
Ascite	0,364	< 0,01	63
Bilirubinémie	0,303	< 0,01	70
Temps de Quick	0,286	< 0,01	67
Cause de l'hémorragie	0,158	< 0,05	61
Médicaments gastro-agressifs	0,154	< 0,05	58
Type de l'hémorragie	0,085	ns	48
Etc.			
Analyse discriminante multidimensionnelle			
Covariables	**Coefficient de Mahalanobis**	**% de sujets bien classés**	
Ascite	0,364	63	
Ascite + bilirubinémie	0,587	72	
Ascite + bilirubinémie + cause de l'hémorragie	0,764	72	
Ascite + bilirubinémie + cause + Quick	0,864	75	
17 covariables	1,710	79	

Cet exemple montre encore que, si le coefficient de Mahalanobis augmente en ajoutant des variables dans le modèle, le passage des quatre variables les plus discriminantes aux 17 variables étudiées dans le modèle, le gain en malades bien classés ne passe que de 75 % à 79 %.

Les malfaçons des études multifactorielles

Une malfaçon courante des études multifactorielles est l'inclusion d'un nombre de covariables trop important par rapport à l'effectif de l'échantillon étudié et surtout d'événements. On admet généralement que l'on ne doit pas inclure plus d'une covariable pour dix événements dans l'échantillon étudié. Par exemple, si dans une étude multifactorielle sur le pronostic vital des exérèses pelviennes dans

des cancers très étendus du rectum, il est observé 16 décès et que l'analyse a inclus huit covariables, les résultats risquent de ne pas être généralisables [7]. Il n'aurait été correct pour 16 événements, de n'inclure qu'une, voire deux covariables dans le modèle. Dans une étude cas-témoin, on analysera de même la variable par groupe de 10 cas supplémentaires.

Les autres malfaçons sont les utilisations inadaptées d'un modèle multifactoriel. Ainsi, dans le modèle de Cox, comme nous l'avons indiqué, il doit y avoir un risque instantané constant de la survenue d'événements que l'on cherche à expliquer : c'est l'hypothèse dite des « hasards proportionnels ». Cela signifie que, si la variable expliquée est la survenue d'une récidive de la maladie, la probabilité de survenue d'une récidive doit être la même à chaque instant. Dans les études cliniques, cette hypothèse est généralement admise, mais très rarement vérifiée. Par exemple, si l'on inclut dans une survie la mortalité postopératoire, la probabilité de décès après une intervention chirurgicale importante comportant un risque non négligeable, sera plus élevée en postopératoire immédiat qu'ultérieurement.

Autre exemple, une régression linéaire est parfois une mauvaise représentation de la réalité. La droite de régression produite sera alors un mauvais modèle prédictif.

Une autre malfaçon est l'inclusion dans le modèle de covariables liées entre elles. Par exemple ictère et hyperbilirubinémie ou encore une lymphocytose CD4, exprimée à la fois en pourcentage et en valeur absolue.

Les autres utilités des analyses multifactorielles

Les comparaisons : là où un essai randomisé n'est pas possible.

Nous avons vu que l'essai randomisé était la méthode qui permettait, dans une comparaison, de se donner le plus de chances, de comparer, au sein de l'ensemble de la population étudiée, deux sous-groupes similaires.

Il est cependant des questions qu'il n'est pas possible de résoudre, pour des raisons techniques ou éthiques à l'aide d'un essai randomisé. L'analyse multifactorielle représente alors la méthode qui se rapproche le plus d'un essai randomisé, bien que l'abord méthodologique soit complètement différent. Nous en donnerons un exemple dans la cinquième partie consacrée au traitement.

Utilisation prédictive des analyses multifactorielles

Un autre intérêt des études multifactorielles prédictives est l'élaboration de scores prédictifs.

Le principe en est le suivant : les analyses multifactorielles permettent, nous l'avons vu dans des exemples, d'estimer les liens qui existent entre des variables expliquantes et la variable expliquée, ainsi que la force de ces liens exprimée en termes d'*odds ratio*, de risques relatifs, de coefficient de régression, voire de coefficient de Mahanalobis dans une analyse discriminante.

À partir de la force de ces liens, il est possible d'affecter à chaque covariable retenue en analyse multifactorielle, un coefficient. Celui-ci est calculé à partir du coefficient de régression. Dans un second temps, il est possible de calculer pour chaque malade inclus dans l'étude son score qui est la somme des scores des covariables statiquement significatives présentes chez lui. Enfin, des groupes de malades peuvent être déterminés en fonction de groupes de scores. En quelque sorte, la régression multiple, la régression logistique ou le modèle de Cox peuvent atteindre le même objectif que celui des analyses discriminantes.

Nous donnerons quelques exemples de l'utilité de ces scores à visée prédictive à propos de la démarche diagnostique et pronostique (quatrième et sixième parties).

Ces scores posent cependant plusieurs problèmes :

1. Comme le montre l'exemple des hémorragies digestives et du coefficient de Mahanalobis, plus les covariables sont nombreuses, plus la valeur prédictive du score est élevée, mais plus il est compliqué et, de ce fait, moins il a de chance d'être utilisé dans la pratique médicale. Il y a donc un choix à faire entre exhaustivité complexe et simplification utile.
2. Ces scores ayant été déterminés sur des échantillons donnés, ils ne peuvent être extrapolés que sur des populations qui répondent très précisément aux critères d'inclusion des sujets qui ont servi à les déterminer. Malgré cela, des biais sont toujours possibles. Il est donc souhaitable de valider les scores qui sont élaborés. Pour ce faire, deux ordres de méthode peuvent être utilisés. Le premier consiste à valider le score proposé sur un ou plusieurs échantillons autres que celui qui a permis son élaboration. C'est la validation externe. Le second ordre de méthode consiste à faire, au sein de l'échantillon étudié, des validations croisées [8].

2

Grille de lecture (ou de réalisation) d'une étude multifactorielle

1. L'objectif de l'étude est pertinent sur le plan médical, c'est-à-dire que :
- le choix de la population étudiée ne doit pas aboutir à confirmer ce qui est déjà bien établi ;
- les covariables étudiées potentiellement importantes ont bien été incluses dans le modèle.

2. Les définitions fondamentales sont précises :
- population étudiée ;
- covariables expliquantes ;
- variable expliquée.

3. Le choix du modèle est correct :
- si la variable expliquée est quantitative, régression multiple ;
- si la variable expliquée est qualitative à deux classes, régression logistique ;
- si la variable expliquée est censurée, modèle de Cox ;
- analyse discriminante dans certains cas.

4. L'analyse a comporté d'abord une étude unifactorielle pour sélectionner ($p < 0,20$ ou $p < 0,25$) les covariables retenues dans l'analyse multifactorielle ;
- il n'a pas été retenu plus d'une covariable par dix événements, deux pour 20, trois pour 30, etc.

5. L'idéal : les résultats ont été validés sur un échantillon différent de celui qui a servi à établir le modèle ou par des méthodes particulières.

Références

1. Cosnes J, Gendre JP, Lacaine F, Naveau S, Le Quintrec Y (1982) Rôles compensateurs de l'iléon et du côlon, restant après résection étendue de l'intestin grêle. Gastroenterol Clin Biol 6: 159-65
2. Rullier E, Laurent C, Garrelon JL *et al.* (1998) Risk factors for anastomotic leakage after resection of rectal cancer. Br J Surg 85: 355-8
3. Cox DR (1972) Regression models and life-tables (with discussion). J R Statis Soc Br 34: 187-220
4. Petrequin P, Huguier M, Lacaine F, Houry S (1997) Cancers de l'œsophage réséqués : modèle prédictif de survie. Gastroenterol Clin Biol 21: 12-6
5. Mahalanobis PC (1936) On the generalised distance in statistic. Proc Ntle Institute Science India 2: 49-55
6. Poynard T, Chaput JC, Mary JY, *et al.* (1980) Analyse critique des facteurs liés à la mortalité au trentième jour dans les hémorragies digestives hautes du cirrhotique. Gastroenterol Clin Biol 4: 655-65
7. Birkmeyer JD, Finlayson SR (1998) Misuse of multivariate analysis. Surgery 124: 114
8. Barrier A, Boelle PY, Lemoine A, *et al.* (2007) Génomique somatique et pronostic des cancers colorectaux. Bull Acad Ntle Med 191:1091-103

La causalité

3

La question d'une relation de nature causale entre un facteur de risque et une maladie est un sujet qui est souvent abordé, notamment par les médias, avec une légèreté inversement proportionnelle à la difficulté de la réponse scientifique à la question, qu'il s'agisse de facteurs environnementaux ou médicamenteux. En effet :

– une causalité est souvent difficile à prouver ;
– il peut y avoir une relation de nature causale entre un facteur de risque et une maladie sans que la responsabilité de la survenue de cette maladie soit toujours attribuable à ce facteur. Ainsi, le tabagisme est une cause de survenue d'un cancer du poumon, mais des cancers du poumon peuvent survenir en dehors de tout tabagisme. Autrement dit, l'existence d'une relation de causalité bien établie permet seulement d'affirmer que la probabilité de développer une maladie liée à ce facteur de risque est plus élevée chez les personnes exposées à ce facteur de risque que chez les autres ;
– il convient enfin de tenir compte du risque en excès et du risque attribuable, c'est-à-dire des mesures d'impact.

Les mesures d'impact

Le risque relatif mesure, on l'a vu, les conséquences individuelles de la présence ou non d'un facteur de risque. Concernant l'individu, il influence la décision médicale pour un sujet donné.

Les mesures d'impact ont un objectif différent. Elles mesurent les conséquences de l'exposition à un facteur de risque en santé publique.

Le risque en excès

Le risque en excès représente la différence entre le risque de survenue d'une maladie chez les sujets exposés au risque et ceux qui ne

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

le sont pas. Il est le résultat d'une soustraction et non d'une division comme le risque relatif, ce dont on se rend compte en comparant le tableau I tiré des mêmes données que le tableau III du chapitre précédent (cf. page ***).

Tableau I – Risque en excès (à partir des mêmes données que le tableau IV de la p. 128.).

Exemple fictif : effet d'une chimiothérapie sur la mortalité dans un cancer			
	Patients décédés	**Patients vivants**	***Total***
Chimiothérapie	63	39	*102*
Pas de chimiothérapie	70	34	*104*
Total	*133*	*73*	*206*
Le risque en excès est égal à la différence entre le risque de mortalité chez les patients qui ont eu de la chimiothérapie (63/102) moins le risque chez les patients n'ayant pas eu de chimiothérapie (70/104) soit : (63/102) – (70/104) = – 0,056			
De façon plus générale, si les données sont les suivantes :			
	Malades	**Non malades**	***Total***
Exposés au risque (E)	*a*	*b*	*l*1
Non exposés (E -)	*c*	*d*	*l*2
Total	*c*1	*c*2	*N*
Le risque en excès est égal à : (*a*/*l*1) – (*c*/*l*2).			

Dans cet exemple, la chimiothérapie diminue donc le risque de décès. Si l'on inversait la proposition, ce qui serait peut-être plus facile à comprendre, le risque de mortalité en excès chez les malades qui n'ont pas de chimiothérapie serait de (70/104) – (63/102) = 0,056. C'est la partie du risque absolu qui semble due au facteur de risque.

Le risque attribuable

Le risque attribuable est encore dénommé « fraction étiologique » du risque. Il permet des décisions en santé publique. Par exemple, si 8 % à 12 % des insuffisances rénales terminales sont attribuables à la prise de fortes doses cumulées de paracétamol [1], cela signifie que si ce médicament était retiré du marché on pourrait éviter, au maximum, 8 % à 12 % des insuffisances rénales terminales.

Il y a deux formulations du risque attribuable. Celui chez les sujets exposés et celui attribuable en population. Dans le premier cas, le risque attribuable mesure la fraction, dans notre exemple, d'insuffisances rénales terminales dues au paracétamol chez les consommateurs de ce produit. Ce risque n'est pas de 100 % parce qu'un consommateur de paracétamol peut développer une insuffisance rénale due à une autre cause. Ce risque attribuable est égal au risque relatif (RR) – 1, divisé par le risque relatif :

$$\text{Risque attribuable} = \frac{\text{Risque relatif} - 1}{\text{Risque relatif}}$$

Toujours dans notre exemple, une étude ayant montré que chez les personnes qui ont absorbé au cours de leur vie entre 1 000 et 5 000 comprimés de paracétamol, ce qui représente 11,9 % de la population (les Français sont les plus gros consommateurs de médicaments au monde avec les Nord-Américains), le risque relatif associé à l'insuffisance rénale terminale était égal à 2,0 [2]. Le risque attribuable chez les consommateurs de paracétamol était alors égal à : (2,0 – 1)/2,0 = 50 %.

L'autre formulation du risque attribuable est le risque attribuable en population. C'est, par exemple, la fraction des insuffisances rénales attribuables au paracétamol dans la population générale qui est de 8 % à 12 % (tableau II). Ce risque attribuable en population est égal à 9,2 %.

Tableau II – Risque attribuable en population (RAP).

$\text{RAP} = \frac{P \text{ exposés} \cdot (\text{RR} - 1)}{1 + P \text{ exposés} \cdot (\text{RR} - 1)}$	
	Dans notre exemple :
P exposés est la proportion de sujets exposés dans la population	11,9 %
RR est le risque relatif	2,0
RAP =	9,2 %

Interprétation

La notion de risque attribuable doit être interprétée avec beaucoup de discernement. En effet, ce risque étant calculé à partir du risque relatif, il est soumis aux biais qui peuvent affecter ce dernier. Dans l'exemple du paracétamol, une mauvaise approximation du risque relatif sur un échantillon isolé et peu représentatif, aurait pu montrer des

valeurs très différentes des 8 % à 12 % observés dans différentes études publiées. Il est alors facile de concevoir qu'un résultat spectaculaire, obtenu à partir d'un échantillon pour lequel il existe un biais, soit attractif pour les médias qui se préoccupent assez peu de la rigueur scientifique avec laquelle les résultats ont été obtenus. Un exemple caricatural est l'alarme récurrente entre la pollution atmosphérique qui serait responsable directement ou indirectement de milliers de décès en France. Si ce risque attribuable était estimé à partir du risque relatif, il conviendrait d'être certain que les études qui ont estimé ce risque relatif de l'association pollution athmosphérique-décès ont tenu compte de nombreux facteurs de confusion : tabagisme, âge, profession, etc.

La causalité

Le problème essentiel de l'épidémiologiste est de déterminer si une différence estimée entre exposés et non exposés n'est qu'une association non causale ou bien s'il existe un effet causal. Cette notion de causalité est de nature probabiliste. Tous les malades qui ont un mésothéliome pleural n'ont pas été en contact avec l'amiante et toutes les personnes ayant été en contact avec l'amiante n'auront pas de mésothéliome. En revanche, en moyenne, une personne en contact avec l'amiante a plus de risque d'avoir un mésothéliome. L'interprétation des faits est souvent plus compliquée encore. Par exemple, on observe que la consommation d'alcool est plus élevée chez les fumeurs que chez les non-fumeurs, ce qui pourra entraîner une plus forte incidence des cancers du poumon chez les alcooliques (association non causale) alors que c'est le tabac qui en est responsable (association causale).

Pour prouver qu'un facteur de risque est non seulement associé à la survenue d'une maladie, mais encore responsable, une accumulation d'arguments est nécessaire [3]. La causalité ne peut pas être établie simplement sur des critères statistiques. Notamment, on a vu l'existence de facteurs de confusion. Ceux-ci sont des obstacles à l'interprétation causale. Le problème est que l'on n'est jamais certain d'avoir observé tous les facteurs de confusion possibles.

Pour les médicaments mis sur le marché, l'alerte au départ peut être donnée par les médecins et les pharmaciens, et depuis juin 2011, par les patients soit à la Commission nationale de pharmacovigilance, soit à la base européenne qui centralise les informations de tous les pays membres. L'expérience de la vaccination contre la grippe H5N1 dans laquelle cette possibilité a été ouverte aux patients a montré

que, sur le total des effets indésirables signalés, 20 % l'avaient été par eux. En pratique, c'est d'abord un Comité technique de pharmacovigilance qui examine, juge les effets indésirables qui ont été signalés avant d'émettre un avis, lui-même transmis à la Commission. Ensuite, un traitement biostatistique, épidémiologique est, au minimum, nécessaire. Cependant, les notifications spontanées ne permettent que rarement une collecte exhaustive de l'ensemble des cas survenus en raison d'une sous-notification habituelle des effets indésirables. D'autres méthodes sont souvent nécessaires : suivis de cohorte, études de cas-témoins, etc. Elles demandent parfois de longues durées d'observation et sont onéreuses. Ce type d'études a ainsi montré qu'il existait une relation entre la prise de certains médicaments comme les benzodiazépines, les antidépresseurs ou les dérivés nitrés et la survenue de chutes.

De façon plus générale, la **détermination d'une causalité implique** :

– de montrer que le risque de maladie est plus élevé lorsque l'on est exposé au facteur de risque considéré qu'en cas contraire. C'est un des objectifs des études épidémiologiques. Comme nous l'avons vu, la force de ces associations est estimée par des mesures statistiques, dont le risque relatif (cf. p. 128). Par exemple, le Centre international de recherche sur le cancer (CIRC), agence de l'Organisation mondiale de la santé, est chargé de dresser la liste des agents qui peuvent être considérés comme cancérogènes pour l'homme.
– D'autres arguments sont la stabilité de l'association dans des recherches différentes, autrement dit, à partir d'associations déjà observées, leur validation sur des populations différentes.
– L'existence d'une relation dose-réponse, c'est-à-dire l'observation que, plus le facteur de risque est important en dose et/ou en durée, plus le risque augmente.
– Bien entendu, l'exposition doit précéder l'apparition de la maladie.
– Il peut s'y ajouter des apports de plausibilité biologique, physiopathologique et des arguments expérimentaux.

Néanmoins, tous ces arguments ne doivent pas être nécessairement présents. Par exemple, l'allergie n'est pas dose-dépendante et sa reproductibilité n'est pas constante. En définitive, aucun de ces arguments ne peut apporter une preuve indiscutable de la causalité et aucun ne doit être considéré comme un critère indispensable pour affirmer la causalité. Cela explique que le cheminement soit long pour accumuler les observations qui permettent de déterminer une forte probabilité de causalité entre un facteur de risque et l'apparition d'une maladie.

Références

1. Ronco PM, Flahault A (1994) Drug-induced end-stage renal disease. N Engl J Med 334: 1711-2
2. Perneger TV, Whelton PK, Klag MJ (1994) Risk of kidney failure associated with the use of acetaminophen, aspirin, and nonsteroidal antiinflammatory drugs. N Engl J Med 331: 1675-9
3. Flahault A, Spira A (2011) La situation épidémiologique en France en 2011. Rapport. Bull Acad Ntle Med (sous presse)

Partie

Le diagnostic

4

Introduction

Les éléments sur lesquels le médecin s'appuie pour faire le diagnostic d'une maladie sont des symptômes (ou signes fonctionnels), des données d'examen (ou signes physiques), des antécédents et, le cas échéant, des examens complémentaires. Ces derniers peuvent être biologiques, radiologiques, isotopiques, etc. Chacune de ces catégories d'examens complémentaires comprend elle-même une diversité d'explorations de plus en plus nombreuses.

L'appréciation de la valeur diagnostique de ces symptômes, signes, examens complémentaires s'est longtemps faite de façon assez subjective en fonction de leur valeur intrinsèque, mais aussi de l'expérience du médecin, voire de préférences subjectives de chacun. Ainsi, il y a une trentaine d'années, à propos de la cholécystite aiguë, il était écrit dans un traité que l'ictère était fréquent, dans un deuxième qu'il était à rechercher, dans deux autres qu'il se voyait dans 10 % des cas, et n'était pas mentionné dans deux autres [1].

De façon plus générale, la valeur d'un élément diagnostique était souvent qualifiée de façon subjective : signe « fréquent » ou « bon examen » ou, à l'inverse, examen « peu fiable », etc.

Ces appréciations sont aujourd'hui devenues objectives et peuvent être quantifiées avec des outils de mesure aussi précis que le centimètre ou la balance. Ils constituent autant d'aides à la décision dans laquelle les contreparties des examens en termes de désagrément et de risque pour les malades, grâce aux progrès technologiques, sont de plus en plus réduites. En revanche, leurs coûts sont de plus en plus élevés. La plus grande partie est prise en charge par la solidarité nationale. Ainsi, le médecin prescripteur d'examens complémentaires a une double responsabilité : scientifique et économique.

Référence

1. Languille T, Flamant Y, Maillard JN (1980) La douleur biliaire aiguë. Essai de sémiologie critique. Gastroentérol Clin Biol 4: 844-7

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Les « outils » de mesure 1

Au sein d'une population bien définie, l'examen que l'on cherche à évaluer peut être anormal ($S+$) ou normal ($S-$). Le critère de jugement est, soit la présence d'une maladie ($M+$), soit son absence ($M-$). Il est alors possible de déterminer, au sein de la population étudiée quatre sous-groupes comme le montre le tableau I.
Insistons, à nouveau, sur la nécessité de définir clairement au départ, d'une part les critères sur lesquels l'examen a été considéré comme normal ou anormal, d'autre part les critères sur lesquels on a mis en évidence la présence ou l'absence de la maladie, appelés le **référentiel externe** (*gold-standard* en anglais) (cf. *infra*).

Tableau I – Les données.

	$M+$	$M-$	**Total**
$S+$	a	b	$a+b$
$S-$	c	d	$c+d$
Total	$a+c$	$b+d$	N

$M+$ représente, au sein de la population étudiée, les malades.
$M-$ les personnes qui n'ont pas la maladie.
$S+$ représente la présence du signe (ou le résultat anormal de l'examen).
$S-$ l'absence du signe (ou le résultat normal de l'examen).

Les quatre sous-groupes du champ du tableau correspondent ainsi :
a à l'effectif des signes présents chez les malades ;
b à l'effectif des signes présents chez des sujets qui n'ont pas la maladie (appelés « faux positifs ») ;
c à l'effectif des signes absents chez les malades (appelés « faux négatifs ») ;
d à celui des signes absents chez les sujets qui n'ont pas la maladie.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

1

À partir des effectifs de ces quatre sous-groupes, il est possible de mesurer à l'aide de variables qualitatives la valeur du signe étudié dans la maladie [1]. Ces outils de mesure sont des applications des probabilités conditionnelles et du théorème de Bayes [2].

Sensibilité et spécificité (tableau II)

Tableau II – Sensibilité et spécificité d'un signe.

	$M+$	$M-$	**Total**
$S+$	a	b	$a+b$
$S-$	c	d	$c+d$
Total	$a+c$	$b+d$	N
La sensibilité (*Se*) est égale à $a / (a+c)$. La spécificité (*Sp*) est égale à : $d / (b+d)$.			

La sensibilité

La sensibilité d'un signe est le pourcentage de cas où il est présent chez les malades, ce que l'on peut encore exprimer en disant que c'est la probabilité du signe s'il y a la maladie. Elle est estimée par le rapport du nombre de malades chez lesquels le signe est présent (a) au nombre de malades ($a + c$).

Un signe est d'autant plus sensible qu'il est souvent présent dans la maladie. Si $c = 0$, c'est-à-dire pas de « faux négatifs », la sensibilité du signe est égale à 1 (ou 100 %), ce qui signifie que le signe est toujours présent chez les malades. C'est le cas de la fièvre dans la typhoïde, ou de l'élévation des transaminases dans les hépatites. Beaucoup d'études qui concernent l'évaluation d'un moyen diagnostique sont rétrospectives et ne portent que sur des patients atteints d'une maladie, chez lesquels un examen complémentaire a été étudié. De ce fait, elles ne peuvent apprécier que la sensibilité de cet examen, ce qui en limite beaucoup l'intérêt comme nous allons le voir.

La spécificité

La spécificité d'un signe dans une population mesure le pourcentage de sujets chez lesquels il est absent parmi ceux qui n'ont pas la maladie. En termes de probabilité, la spécificité estime la probabilité

de l'absence du signe en l'absence de maladie. La spécificité est, en effet, estimée par le rapport du nombre de sujets qui n'ont pas la maladie ni le signe (*d*), à l'ensemble des sujets qui n'ont pas la maladie dans la population étudiée (*b* + *d*).
Un signe est d'autant plus spécifique qu'il est rarement présent chez les personnes qui n'ont pas la maladie. S'il n'y a pas de « faux positifs » (*b* = 0), la spécificité du signe est égale à 1 (ou 100 %). C'est ce que l'on appelle alors un signe pathognomonique de la maladie comme le signe de Koplik dans la rougeole. Ce type de signe est malheureusement très rare.

Les limites de la sensibilité et de la spécificité

Pour intéressants que soient les estimations de la sensibilité et de la spécificité, que l'on peut assortir de leur intervalle de confiance, ces deux moyens de mesure ne permettent pas de répondre aux deux principales questions qui intéressent le clinicien devant la présence d'un signe clinique ou devant le résultat d'un examen complémentaire : si le signe est présent, quelle est la probabilité que le sujet ait la maladie que l'on cherche à diagnostiquer ? Et si le signe est absent, quelle est la probabilité que le sujet n'ait pas la maladie que l'on a pu évoquer ?
Les valeurs prédictives répondent à ces interrogations. Ces valeurs, encore dénommées probabilités *a posteriori*, sont, comme la sensibilité et la spécificité, des probabilités conditionnelles dont les développements mathématiques ont été formulés, eux aussi par Thomas Bayes[1] (cf. *infra*).

Valeurs prédictives (tableau III)

Tableau III – Les valeurs prédictives.

	M +	*M* –	**Total**
S +	*a*	*b*	*a* + *b*
S –	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*N*
La valeur prédictive positive (VPP) est égale à *a*/*a* + *b*. La valeur prédictive négative (VPN) est égale à : *d*/*c* + *d*.			

1 Bayes était un pasteur anglican ayant vécu au XVIII[e] siècle.

1

La valeur prédictive positive (VPP)

La valeur prédictive positive d'un signe estime la probabilité de la maladie chez les personnes qui ont ce signe. Cette estimation correspond au rapport du nombre de sujets qui ont la maladie et chez lesquels le signe est présent (a) sur le nombre de sujets chez lesquels le signe est présent ($a + b$).
La VPP d'un signe est d'autant plus grande que le signe est rarement présent chez les personnes qui n'ont pas la maladie, autrement dit que le nombre de « faux positifs » est faible.

La valeur prédictive négative (VPN)

La valeur prédictive négative d'un signe estime la probabilité d'absence de la maladie chez les personnes qui n'ont pas ce signe. Elle est définie par le rapport du nombre de sujets qui n'ont pas la maladie et chez lesquels le signe est absent (d) sur le nombre de sujets chez lesquels le signe est absent ($c + d$).
La VPN d'un signe est d'autant plus grande que l'absence du signe est rare chez les personnes qui ont la maladie, autrement dit que le nombre de « faux négatifs » est faible.

Le lien entre ces quantités

Le théorème de Bayes permet, de façon générale, d'estimer la probabilité de survenue d'un événement, sachant qu'un autre événement est connu (tableau IV). Par exemple, il estime la probabilité d'une maladie lorsqu'un signe pathologique est présent en tenant compte de la prévalence de la maladie dans la population étudiée.

Tableau IV – Valeurs prédictives et caractéristiques du test.

Le théorème de Bayes permet d'écrire :

$$p(M+|S+) = \frac{p(M+) \times p(S+|M+)}{p(M+) \times p(S+|M+) + p(M-) \times p(S+|M-)}$$ dans lequel :

$p\ (M+)$ est la prévalence P de la maladie dans la population étudiée.

$p\ (M-)$ est le complément de la prévalence $1 - p\ (M+)$

| Cette barre verticale exprime une probabilité conditionnelle :

$p\ (S+\ |M+)$ est la probabilité du signe si la maladie est présente, c'est-à-dire la sensibilité (Se) du signe.

$p\ (S+\ |\ M-)$ est la probabilité du signe s'il n'y a pas la maladie ; c'est le complément de la spécificité c'est-à-dire $1 - Sp$.

Appliqué aux valeurs prédictives positives (VPP) et négatives (VPN), le théorème de Bayes peut donc s'écrire :

$$\text{VPP} = p(M+|S+) = \frac{P \times Se}{P \times Se + (1-P) \times (1-Sp)}$$

De façon analogue, la valeur prédictive négative, c'est-à-dire la probabilité d'absence de la maladie si le signe est absent, s'écrit :

$$\text{VPN} = p(M-|S-) = \frac{(1-P) \times Sp}{(1-P) \times Sp + (1-p) \times (1-Se)}$$

Il est possible, par analogie, de calculer de la même façon la probabilité de maladie si le signe est absent ou la probabilité d'absence de maladie si le signe est présent, mais ces probabilités sont, en général, moins utiles en pratique médicale.

Question préalable en guise d'exercice

On avait donné aux médecins de la célèbre Harvard Medical School de Boston les trois informations suivantes [3] :

- la fréquence des hépatites, présumées virales, dans une population générale nord-américaine est de 1 pour 1 000 ;
- les sujets qui ont une hépatite virale ont toujours une élévation des transaminases ;
- mais, dans la population générale, on observe que 5 % des sujets peuvent avoir une élévation des transaminases sans avoir pour autant une hépatite virale.

Il a été ensuite posé à ces médecins la question suivante : si vous voyez une personne qui a une élévation des transaminases, quelle est la probabilité qu'il ait une hépatite virale ?

Seulement 18 % des médecins interrogés ont su répondre correctement à cette question. Nous vous proposons de tenter dès maintenant l'exercice. Nous vous assurons que la lecture des paragraphes suivants vous permettra d'y arriver. Vous pourrez confronter votre réponse à la solution qui sera donnée ensuite (tableau V).

Réponse à la question posée aux médecins de Harvard

Il est possible de répondre en remplissant pas à pas le tableau à quatre cases (tableau I) en reprenant l'énoncé des données et en déduisant certaines données :

Tableau V – Réponse à l'exercice.

1) La fréquence des hépatites présumées virales dans une population générale nord-américaine est de 1 pour 1 000, ce qui donne :

	M +	*M* –	**Total**
S +	*a*	*b*	*a* + *b*
S –	*c*	*d*	*c* + *d*
Total	1	**999**	**1 000**

2) Les sujets qui ont une hépatite virale ont toujours une élévation des transaminases, ce qui donne :

	M +	*M* –	**Total**
S +	**1**	*b*	*a* + *b*
S –	0	*d*	*c* + *d*
Total	**1**	**999**	**1 000**

3) Dans la population générale, on observe que 5 % des sujets peuvent avoir une élévation des transaminases sans avoir pour autant une hépatique virale, ce qui donne (par approximation) :

	M +	*M* –	**Total**
S +	**1**	**50**	**51**
S –	0	**949**	**949**
Total	**1**	**999**	**1 000**

La sensibilité (*Se*) du signe est 1/1 soit 100 % ou 1,0 ; la spécificité (*Sp*) est 949/999 ou 0,05.

$p\,(M+) = 0{,}001.$

$p\,(S+\,|M+\,) = Se = 1{,}0$

$p\,(S+\,|M–) = Sp = 0{,}05$

$$\text{La VPP } = p(M+|S+) = \frac{0{,}001 \times 1{,}0}{0{,}001 \times 1{,}0 + (1-999) \times (1-0{,}05)} = 0{,}02, \text{ soit 2 \%}$$

Il est encore plus simple d'appliquer la formule du tableau IV qui donne :

VPP = 1/51, soit approximativement 2 %, et

VPN = 949/949, soit 100 %.

Références

1. Lacaine F, Huguier M, Gremy F (1978) L'efficacité d'un examen à but diagnostique : de la donnée à la décision médicale. Nouv Presse Med 7: 1451-3
2. Price (1763) An essay towards solving a problem in the doctrine of chances by the late Rev. Mr Bayes, F.R.S. Philosophical transactions: 370-418. (Nous remercions la bibliothécaire de la faculté de médecine de Lille qui nous a adressé une photocopie de la publication originale de la communication de Price)
3. Casscells W, Schoenberger A, Graybos T (1978) Interpretation by physicians of clinical laboratory results. N Engl J Med 299: 999-1000

Remarques sur la sensibilité, la spécificité, les valeurs prédictives. Les courbes ROC

2

Les trois grandes définitions

L'évaluation de la valeur diagnostique d'un symptôme, d'un signe ou d'un examen complémentaire dans une maladie dépend des définitions :
– de la population étudiée dans laquelle il y a des malades, mais aussi des non-malades ;
– des critères sur lesquels on a déterminé que l'examen était normal ou anormal ;
– des arguments sur lesquels on a déterminé que la maladie était bien présente ou, dans le cas contraire absente ; c'est, nous l'avons indiqué, le référentiel ou le standard de référence externe.

La définition de la population (ou de l'échantillon) inclus dans l'étude

L'importance de cette définition dans l'interprétation des résultats va être montrée en prenant comme exemple l'échographie transcutanée dans le diagnostic de métastases hépatiques chez des malades qui ont un cancer primitif connu. Rappelons que le diagnostic de l'existence ou non de ces métastases est fondamental pour orienter la stratégie thérapeutique.

Une étude, réalisée dans un service de radiologie nord-américain, a concerné 189 malades. Elle avait surtout porté sur des malades qui avaient un cancer du sein. Elle a montré que la sensibilité de l'examen était de 82 % [1].

Une autre étude a été menée dans un service de chirurgie sur 273 malades atteints de cancers de l'appareil digestif. La sensibilité de l'échographie avait été de 66 %[2]. La différence entre les résultats des deux études était statistiquement significative.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Plusieurs interprétations ont été évoquées. Les différences pouvaient être dues au fait que les appareils d'échographie étaient sensiblement différents, mais celui de la seconde étude, un peu plus récent que celui utilisé dans la première étude était plutôt plus performant que l'autre. Il pouvait encore s'agir de différence de performance des radiologues américains et français. Une autre explication possible était que, dans un cas, il s'agissait surtout de métastases de cancers du sein et, dans l'autre, surtout de métastases de cancer du côlon qui pouvaient avoir une échogénéicité moindre. En fait, la principale explication était que les malades avaient été inclus sur des critères assez différents d'une étude à l'autre. Dans le travail du service de chirurgie, seuls les patients qui n'avaient pas de métastases cliniquement décelables à la palpation avaient été inclus. Dans l'étude des radiologues, de tels malades étaient inclus. Dans le premier cas, l'examen complémentaire était un examen de dépistage. Dans l'autre cas, de grosses métastases palpables augmentaient, bien entendu, la sensibilité de l'échographie qui ne faisait que les confirmer. Il y avait donc à la base une différence entre les personnes dites « malades » dans les deux études, avec des patients plus atteints dans le second cas. On appelle ce phénomène le **biais de spectre** (*spectrum bias* en anglais).

Ainsi, des différences de sensibilité ou de spécificité d'un examen d'une étude à l'autre peuvent s'expliquer par des différences de populations incluses, c'est-à-dire d'échantillons. Cela montre encore qu'un travail mené avec une bonne rigueur méthodologique est d'autant plus intéressant qu'il est plus pertinent : l'évaluation de la sensibilité de l'échographie, chez un malade qui a un cancer connu et une métastase hépatique palpable cliniquement, a un intérêt plus limité que si le foie paraît normal à la palpation.

Sur quels critères l'examen que l'on cherche à évaluer a-t-il été considéré comme positif (anormal) ou négatif (normal) ?

Le choix des critères de normalité ou d'anormalité n'est pas toujours évident. Il affecte cependant les valeurs de la sensibilité, de la spécificité ou des valeurs prédictives d'un examen. Dans notre exemple de l'échographie dans le diagnostic de métastases hépatiques, il convient dans le protocole d'étude de définir les critères de diagnostic de métastases. En effet, toute tumeur hépatique n'est pas forcément une métastase, mais peut être un adénome, une hyperplasie nodulaire focale, etc. Pour un examen biologique, il est possible de prendre comme limite supérieure de la normale ou bien celle indiquée par le laboratoire ou bien la valeur observée, chez les malades inclus dans l'étude et qui n'ont pas

la maladie, plus ou moins deux écarts-types. Ainsi, deux études ont évalué la sensibilité et la spécificité des lactico-déshydrogénases dans le dépistage de métastases hépatiques chez des malades qui avaient un cancer colorectal connu. Le tableau I montre les résultats de ces deux études.

Tableau I – Résultats de deux études sur la sensibilité et la spécificité des lactico-déshydrogénases dans le diagnostic de métastases hépatiques d'un cancer colorectal.

	Sensibilité (%)	**Spécificité (%)**
Étude de Adloff *et al.* [3]	87	72
Étude de Molkhou *et al.* [2]	51	84

Ces différences dans des examens biologiques réalisés avec la même technique de dosage pouvaient s'expliquer par le fait que dans l'étude d'Adloff *et al.* la limite supérieure de la normale, qui avait été retenue, était celle indiquée par le laboratoire, soit 120 U/L. Dans l'autre étude, les auteurs avaient pris comme limite supérieure de la normale la moyenne de la valeur observée chez les malades qui n'avaient pas de métastases hépatiques plus deux écarts-types, et qui était de 218 U/L. Cet exemple montre encore que, pour un même examen, la sensibilité et la spécificité évoluent toujours de manière antagoniste : renforcer l'une implique que l'on réduise l'autre.

Sur quels critères le diagnostic de maladie ou de non-maladie a-t-il été établi (le référentiel ou standard de référence externe) ?

En continuant à prendre notre exemple du diagnostic de métastases hépatiques, il convient de savoir sur quels critères le diagnostic de métastases a été porté ou récusé. Le critère macroscopique à l'échographie est insuffisant. En effet, chez un malade qui a un cancer colorectal et une lésion hépatique, celle-ci n'est pas une métastase une fois sur cinq [4]. L'examen anatomopathologique est, dans ce cas, le standard de référence externe pour définir l'existence de métastase. En revanche, ce standard est inapproprié pour définir l'absence de métastase et même à l'intervention chirurgicale éventuelle, la palpation du foie peut méconnaître des métastases, notamment centro-hépatiques [5]. L'échographie peropératoire et le suivi du malade sont alors nécessaires pour s'assurer de l'absence de métastases et peuvent, de ce fait, constituer le standard de référence externe de non-maladie [6].

Les courbes ROC

Le signe ou l'examen idéal serait celui qui aurait une sensibilité de 100 % et une spécificité de 100 % et donc des valeurs prédictives, elles aussi, de 100 %. Malheureusement, il n'a pas encore été trouvé. En pratique, il faut faire un compromis entre la sensibilité et la spécificité qui varient en sens opposé. Si l'on privilégie la sensibilité d'un examen, il sera souvent peu spécifique et réciproquement. Le *melæna* est très spécifique d'une hémorragie digestive, mais il est peu sensible : des hémorragies digestives peu abondantes n'entraînent pas de melæna. Inversement, l'Hémocult® est plus sensible, pour dépister du sang dans les selles, mais il est peu spécifique d'hémorragie digestive : un petit saignement d'origine gingivale ou des facteurs alimentaires, par exemple, peuvent rendre un Hémocult® positif.

Indices globaux

Pour essayer de concilier les termes de cette alternative, examen sensible, mais peu spécifique ou spécifique, mais peu sensible, des moyens prenant en compte à la fois la sensibilité et la spécificité ont été proposés.

Ainsi, la « fiabilité » d'un examen estime la somme des « vrais positifs » et des « vrais négatifs » sur l'ensemble des cas étudiés $(a + d)/N$. Un autre outil est le rapport de vraisemblance de l'examen. C'est le rapport du pourcentage des « vrais positifs » chez les malades $(a/a + c)$ sur les « faux positifs » chez les sujets qui n'ont pas la maladie » $(b/b + d)$. Par exemple, un rapport de vraisemblance égal à quatre signifie que l'examen est quatre fois plus souvent positif chez les malades que chez ceux qui n'ont pas la maladie. Par analogie, il est encore possible de mesurer le rapport de vraisemblance négatif qu'est le rapport du pourcentage des « faux négatifs » chez les malades $c/(a + c)$ sur les « vrais négatifs » chez les sujets qui n'ont pas la maladie » $d/(b + d)$. Un rapport de 0,5 signifie que l'absence du signe est deux fois moins souvent observée chez les malades que chez les sujets qui n'ont pas la maladie.

Les *Receiver Operating Characteristics curves* ou courbes ROC

Certains signes diagnostiques sont évalués par une mesure, par exemple un dosage, et demandent la définition d'un seuil pour prendre une décision opérationnelle. Prenons l'exemple du dosage des transaminases chez des personnes atteintes d'une hépatite et chez des sujets sains. La distribution du dosage dans l'un et dans l'autre groupe, en supposant leur

distribution normale, peut être décrite par des courbes de Laplace-Gauss (fig. 1). Il est habituel que ces deux courbes se superposent partiellement avec un chevauchement entre les valeurs les plus élevées chez les non-malades et les valeurs les plus basses chez les malades. Si l'on prend une valeur seuil basse, 100 par exemple, il y aura des faux positifs, les sujets sans hépatite qui ont un taux de transaminases supérieur à 100, et très peu de faux négatifs : l'examen sera peu spécifique, mais très sensible. Si l'on prend au contraire une valeur plus élevée, 300, il y aura très peu de faux positifs, mais des faux négatifs représentés par les malades atteints d'hépatite qui ont un taux de transaminases inférieur à 300 : l'examen sera considéré comme très spécifique, mais peu sensible. On comprend bien que lorsque l'on fait varier le curseur de la valeur seuil en deçà de laquelle on considère que le résultat est normal et au-delà de laquelle on le considère comme étant anormal ou pathologique, on fera varier à chaque fois les valeurs estimées de la sensibilité et de la spécificité de l'examen. C'est ce que permettent de quantifier les courbes ROC.

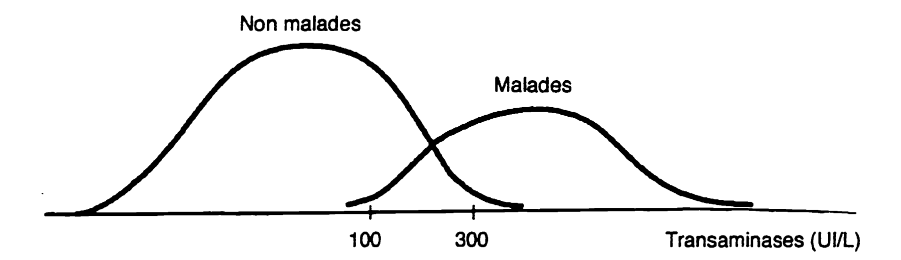


Fig. 1 – Résultats de la mesure d'une variable quantitative dans une population comportant un sous-groupe de malades et un sous-groupe de non-malades. 1) La sensibilité et la spécificité dépendent de la valeur limite de la normale que l'on a choisie. 2) Elles varient en sens opposé : lorsque la sensibilité augmente, la spécificité diminue, et réciproquement.

Les courbes ROC ont été imaginées pendant la Seconde Guerre mondiale par les Britanniques pour régler leurs radars. Le problème était le suivant : les Allemands envoyaient, notamment sur Londres, des missiles chargés d'explosifs, les V1, puis plus tard le V2. Les radars anglais cherchaient à détecter ces missiles dès qu'ils survolaient la Manche afin de les détruire en vol et de déclencher l'alerte pour permettre aux Londoniens de descendre dans les abris. Le problème du réglage des radars s'est alors posé. Si le réglage était trop sensible, un albatros, voire un gros goéland, risquait de déclencher l'alerte. Mais si le réglage était plus spécifique, il risquait d'être insuffisamment sensible et des missiles pouvaient de ne pas être détectés. Des études de sensibilité et de spécificité ont donc été menées pour chercher et pour établir un seuil optimal de détection. C'est le but des courbes ROC.

2

Élaboration des courbes ROC

Les courbes ROC consistent à porter sur un graphique, pour une valeur donnée de seuil entre le normal et l'anormal d'un examen, sa sensibilité (*Se*) en ordonnée et la spécificité correspondante (*Sp*) en abscisse. En répétant ces mesures pour différentes valeurs seuil, on peut déterminer des courbes : c'est le principe des courbes ROC (fig. 2).

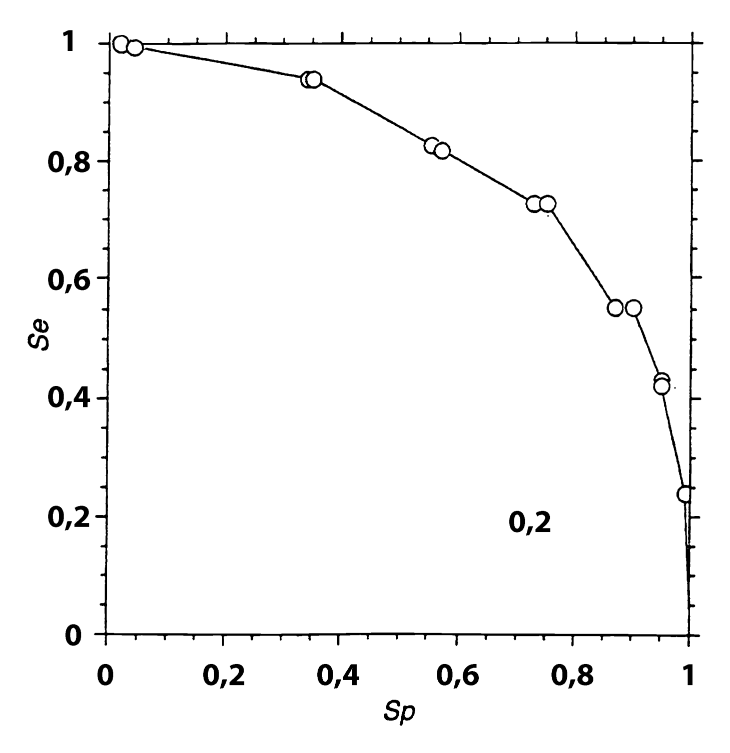


Fig. 2 – Exemple d'une courbe de sensibilité (*Se*) en fonction de la spécificité (*Sp*), mais ceci n'est pas l'expression habituelle d'une courbe ROC (fig. 3).

Ces courbes montrent qu'habituellement, une valeur seuil qui correspondrait à une sensibilité proche de 100 % aurait une spécificité proche de 0 %, alors qu'une valeur seuil proche de 100 % de spécificité aurait une très mauvaise sensibilité. Sur ce graphique, une courbe ROC proche de la diagonale correspond à un test qui n'a pas d'intérêt diagnostique.

En fait, dans les courbes ROC, élaborées à l'aide de logiciels, l'axe des abscisses correspond, non pas à la spécificité, mais à son complément, 1 – *Sp* (fig. 3).

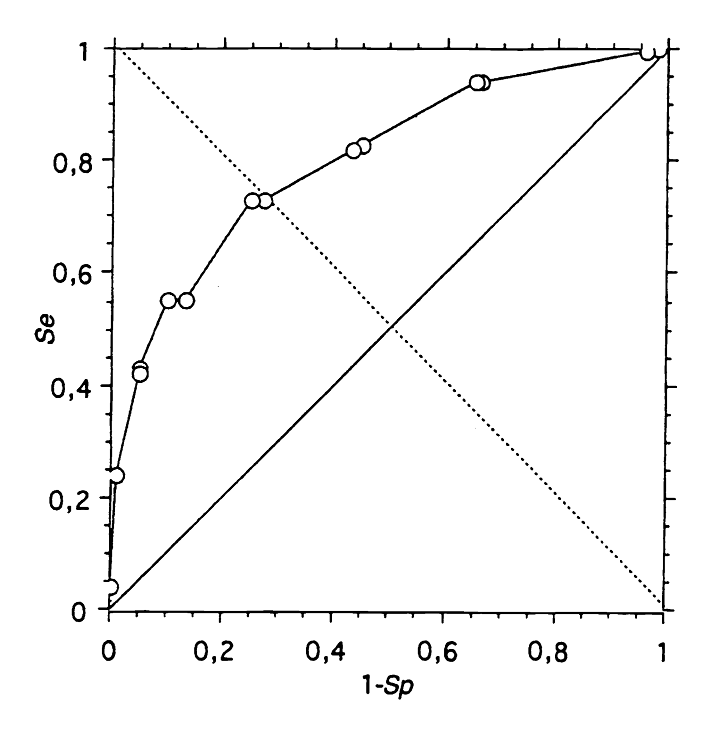


Fig. 3 – Courbe ROC. Par rapport à la figure 2, au lieu de porter en abscisses les valeurs de la spécificité (*Sp*), il est porté leurs compléments (1 – *Sp*).

Sous les conditions rappelées ci-après, la valeur seuil optimale est celle qui correspond sur la courbe à son point d'inflexion, soit encore au point d'intersection de la courbe ROC et de la deuxième bissectrice de l'axe des abscisses et des ordonnées. Ce choix s'applique dans l'hypothèse d'un coût affecté aux erreurs de diagnostic équivalent ou à peu près équivalent pour les « faux positifs » et les « faux négatifs ».

Il existe des tests statistiques qui permettent de comparer les courbes ROC observées à la première bissectrice qui représente une progression linéaire entre la sensibilité et le complément de la spécificité. Il est encore possible de comparer deux courbes ROC entre elles pour deux examens différents. Des tests statistiques reposent sur la comparaison des surfaces sous les courbes (AUROC) (fig. 4). Comme il vient d'être dit, la courbe ROC qui serait sur la première diagonale correspond à un test sans intérêt diagnostique. Dans cette configuration, l'aire sous la courbe AUROC correspondrait à 0,5. La valeur de l'AUROC augmente avec la qualité diagnostique du test dont on peut tester l'intérêt, par exemple, en montrant que son AUROC est significativement plus grande que 0,5.

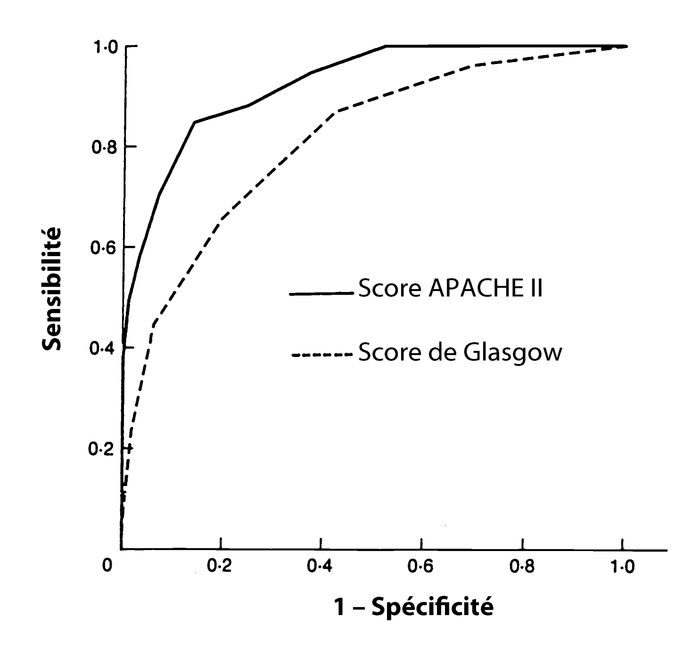


Fig. 4 – Courbes ROC comparant les valeurs diagnostiques de sévérité dans les pancréatites aiguës entre deux scores différents, le score APACHE (aire sous la courbe : 0,78) et le score de Glasgow (aire sous la courbe : 0,65) à l'aide d'un test de McNemar ($p = 0{,}005$) [7].

Rôle de la prévalence de la maladie

La sensibilité et la spécificité sont indépendantes de la prévalence de la maladie dans la population étudiée (bien que des travaux récents aient suggéré une certaine dépendance entre sensibilité, spécificité et prévalence).

En revanche, comme nous l'avons déjà montré à propos du théorème de Bayes, les valeurs prédictives sont étroitement dépendantes de la prévalence de la maladie dans la population étudiée (tableaux III et IV du chapitre I). Le tableau II montre qu'un examen peut avoir la même sensibilité et la même spécificité qu'un autre dans deux échantillons dans lesquels la prévalence de la maladie diffère. Mais les valeurs prédictives diffèrent pour cette même raison.

Ces deux considérations font comprendre que :

– les résultats d'une étude de sensibilité et de spécificité d'un examen peuvent être extrapolés à une autre population que celle à partir de laquelle elles ont été estimées ;

– en revanche, les valeurs prédictives ne peuvent l'être que dans des populations dans lesquelles la prévalence de la maladie est du même ordre que celle de l'échantillon qui a servi à les estimer.

Tableau II – Sensibilité, spécificité, valeurs prédictives et prévalence.

Premier exemple :	*M* +	*M* –	**Total**
S +	*45*	*10*	*55*
S –	*5*	*90*	*95*
Total	*50*	*100*	*150*
Dans ce premier exemple, *Se* = 45/50 = 90 % *VPP* = 45/55 = 82 % *Sp* = 90/100 = 90 % *VPN* = 90/95 = 95% Prévalence = 50/150 = 33 %			
Second exemple :	*M* +	*M* –	**Total**
S +	*18*	*20*	*38*
S –	*2*	*180*	*182*
Total	*20*	*200*	*220*
Dans ce second exemple, *Se* = 18/20 = 90 % *VPP* = 18/38 = 47 % *Sp* = 180/200 = 90 % *VPN* =180/182 = 99 % Prévalence = 20/220 = 10 % Ces deux exemples montrent que, si la sensibilité et la spécificité d'un examen sont indépendantes de la prévalence (et égales dans nos deux exemples), il n'en est pas de même des valeurs prédictives.			

Dans l'exemple du tableau II, les valeurs prédictives estimées sur le premier échantillon, de 82 % pour la VPP et de 95 % pour la VPN, ne peuvent être utilisées dans le deuxième échantillon dans lequel la prévalence de la maladie est de 10 % alors qu'elle était de 33 % dans le premier. Dans le cas contraire, il faudrait recalculer les VPP et les VPN pour la population de patients à laquelle on envisage d'appliquer le test. Pour une sensibilité et une spécificité données, les VPP augmentent avec la prévalence de la maladie dans la population étudiée alors que les VPN diminuent (tableau III).

Tableau III – Exemple théorique d'un examen ayant une sensibilité et une spécificité de 95 % et des valeurs prédictives en fonction de la prévalence de l'affection dans l'échantillon.

Prévalence (%)	**Valeurs prédictives (%)**	
	positives	**négatives**
1	16,1	99,9
2	27,9	99,9
5	50,0	99,7
10	67,9	99,4
20	82,6	98,7
50	95,0	95,0
75	98,3	83,7
100	100,0	–

Effectifs nécessaires pour contrôler la valeur des intervalles de confiance et des indices informationnels d'un examen

Lorsque l'on estime la sensibilité et la spécificité d'un examen, ces mesures doivent être assorties de leur intervalle de confiance. Il est souhaitable que ces intervalles soient aussi réduits que possible. Si un examen a une sensibilité de 80 % et que la mesure de cette sensibilité a été effectuée sur un petit nombre de cas, la valeur inférieure de l'intervalle de confiance à 95 % sera de 60 %, voire de 50 %, ce qui limite considérablement l'intérêt d'une telle étude. Pour réduire cette éventualité, il est possible de calculer les effectifs qu'il est nécessaire d'inclure dans l'étude. C'est ce que montre le tableau IV.

Tableau IV – Exemples du nombre de cas à inclure pour une puissance du test (1 – β) de 95 % en fonction de la sensibilité espérée de l'examen que l'on cherche à évaluer et du seuil de l'intervalle de confiance que l'on se fixe[2].

	Valeur inférieure de l'intervalle de confiance								
	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90
Sensibilité espérée :									
0,60	268	1 058							
0,65	119	262	1 018						
0,70	67	114	248	960					
0,75	42	62	107	230	869				
0,80	28	40	60	98	204	756			
0,85	18	26	33	52	85	176	624		
0,90	13	18	24	31	41	70	235	474	
0,95	11	12	14	16	24	34	50	93	298

Par exemple, si un laboratoire se propose de commercialiser une bandelette urinaire dont il espère que la sensibilité sera de 90 % (0,90) et s'il souhaite que la limite inférieure de l'intervalle de confiance soit 80 % (ou 0,80), il lui faudra inclure 235 échantillons d'urine pour atteindre cet objectif. Si cet examen doit être utilisé dans une population dans laquelle la prévalence de la maladie est de 10 %, il faudra

2 D'après Thomas G. et Flahault A.

qu'il y ait environ 2 100 témoins, c'est-à-dire que l'étude devra porter sur un total de 2 335 échantillons d'urine.

Références

1. Alderson P, Adam DF, McNeil BJ, *et al.* (1983) Computed tomography, ultrasound, and scintigraphy of the liver in patients with colon or breast carcinoma: a prospective comparison. Radiology 149: 224-30
2. Molkhou JM, Lacaine F, Houry S, Huguier M (1989) Dépistage des métastases hépatiques des cancers digestifs. Place des dosages enzymatiques et de l'échographie. Presse Med 18:1370-4
3. Adloff M, Arnaud JP (1985) Étude prospective critique des différentes méthodes de détection des métastases hépatiques. Ann Gastroenterol Hepatol 21: 31-4
4. Chan AW, Altman DG (2005) Identifying outcomes reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 330: 753
5. Smith TJ, Kemeny MM, Sugarbaker PH (1982) A prospective study of hepatic imaging in the detection of metastastic disease. Ann Surg 33: 519-23
6. Raffaelsen SR, Kronborg O, Larsen C, Fenger C (1995) Intraoperative ultrasonography in detection of hepatic metastases for coloretal cancer. Dis Colon Rectum 38: 355-60
7. Mofidi R, Madhavan KK, Garden OJ, Parks RW (2007) An audit of the management of patients with acute pancreatitis against national standards of pratice. Br J Surg 94: 844-8

La démarche diagnostique, choix d'un examen, attitude décisionnelle

3

La démarche diagnostique

Le raisonnement intuitif

La démarche diagnostique intuitive fait appel aux connaissances et à l'expérience. Son analyse permet de comprendre l'aide au diagnostic que peuvent apporter les probabilités bayésiennes.

Cette démarche diagnostique consiste, à partir d'un symptôme pour lequel le malade vient consulter, à évoquer des probabilités diagnostiques *a priori* qui reposent comme il vient d'être indiqué sur les connaissances et l'expérience. Chez un adulte qui se plaint d'une douleur abdominale aiguë, on sait intuitivement qu'il peut avoir une appendicite, une cholécystite ou de nombreuses autres affections moins fréquentes.

Ensuite, toujours intuitivement, on intègre dans le raisonnement d'autres données que le symptôme « douleur abdominale aiguë » (âge, sexe, autres symptômes, signes d'examen, etc.). Ainsi, dans notre exemple, si le malade est jeune, le diagnostic d'appendicite est plus probable que s'il est âgé. La localisation de la douleur dans la fosse iliaque droite augmente encore cette probabilité. En revanche, s'il s'agit d'une femme, la probabilité *a priori* d'une cholécystite est plus élevée que s'il s'agit d'un homme, la lithiase vésiculaire étant plus fréquente chez la femme que chez l'homme, etc.

On aboutit ainsi, à partir des symptômes dont se plaint le malade et de ses antécédents, puis de l'examen clinique à des probabilités diagnostiques *a posteriori*.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Si, au terme de ces étapes du raisonnement, il persiste plusieurs hypothèses diagnostiques, l'étape suivante consiste à demander des examens complémentaires. Nous verrons comment ces derniers peuvent être choisis et comment ils doivent l'être dans une perspective décisionnelle.

Cette démarche intellectuelle fait, à notre avis, une grande partie de l'intérêt de l'exercice médical, ce qui, comme nous allons l'expliquer, n'exclut pas, bien au contraire, l'apport scientifique des probabilités bayésiennes.

La démarche systématique

Malgré sa pratique, nous la déconseillons vivement pour des raisons de raisonnement et de coûts. Elle consiste, à partir d'un ou deux symptômes, à demander toute une série d'examens complémentaires orientés par ce ou ces deux symptômes et à commencer à réfléchir au vu de leurs résultats. Il s'y ajoute habituellement une méconnaissance totale du risque de première espèce inhérent à tous ces examens, ce qui amène à en refaire. Il est possible de comparer cette attitude à celle de joueurs dans les casinos de Las Vegas qui mettent un jeton dans la machine à sous, abaissent la manivelle en espérant qu'ils vont toucher le jack pot. Cependant, la différence est que les médecins prescripteurs ne perdent pas leur argent, mais dépensent celui que la solidarité nationale, par le biais de l'assurance maladie, met indirectement à leur disposition.

« *La prescription d'exploration à la mode, lancée par des travaux de recherche et repris par le grand public médical est une erreur. Les investigations non raisonnées et non guidées par un interrogatoire et un examen clinique correct sont inutiles* », disait déjà en 1977, le professeur Mornex, qui fut doyen et président d'Université [1]. Plus récemment, un enseignant de l'université de Boston aux États-Unis ne dit pas autre chose quand il écrit, en substance [2] : « *Les avancées récentes dans les connaissances scientifiques et la technologie ont permis le développement d'un vaste ensemble de nouveaux tests, de nouveaux agents pharmacologiques, et de traitements. Ils sont si facilement accessibles que peu d'entre nous arrivent à résister à les prescrire à toute occasion […] Ce faisant, nous entraînons la faillite de notre système de santé […] Nous n'acceptons pas l'incertitude […] Il en résulte que, par réflexe, nous surprescrivons des examens et des traitements dans l'idée de protéger nos patients, mais aussi nous-mêmes. Nous croyons que tout faire et le moyen de prévenir le mal et que cela va mettre à l'abri de tout blâme […] Nous devons apprendre à nos étudiants à savoir réfléchir plus, à attendre plus, à observer plus. Nous devons apprendre à nos patients que le plus de*

médecine n'est pas la meilleure médecine et que des prescriptions coûteuses ne sont pas synonymes des meilleurs soins ».

Le raisonnement objectif

À la démarche assez intuitive qui a été décrite au début de ce chapitre, on peut substituer une démarche plus scientifique. Nous allons reprendre l'exemple déjà évoqué d'une douleur abdominale aiguë.
La première étape a consisté à utiliser les résultats des études épidémiologiques sur les causes des douleurs abdominales aiguës chez des patients adultes (plus de 15 ans) venus consulter en urgence à l'hôpital pour ce symptôme. Une étude a ainsi été menée chez 3 500 malades [3]. Elle a montré que les causes en étaient une appendicite dans 26 % des cas, une cholécystite dans 10 % des cas, etc. ou encore un kyste de l'ovaire compliqué dans 2 % des cas. Ce sont les probabilités *a priori* des causes de douleurs abdominales aiguës.
Cette étude a été complétée par la mesure de la sensibilité et la valeur prédictive positive de chaque signe dans chacune des affections qui pouvaient être la cause des douleurs abdominales aiguës.
À partir de ces données, et des probabilités bayésiennes, il a été possible d'élaborer une aide au diagnostic en introduisant dans le modèle, une à une, les données, ce qui aboutit à des probabilités *a posteriori* [4]. Par exemple, la probabilité *a priori* d'une cholécystite était de 10 % comme nous l'avons indiqué. Un âge inférieur à 50 ans diminuait cette probabilité de 10 % à 3 %. L'appartenance au sexe féminin réaugmentait la probabilité de cholécystite de 3 % à 7 %. Le siège de la douleur dans l'hypocondre droit augmentait encore cette probabilité, etc.
Les probabilités bayésiennes peuvent ainsi s'appliquer à plusieurs hypothèses diagnostiques *a priori*. Pour cela, il convient, comme dans notre exemple, que les diagnostics soient mutuellement exclusifs et que toutes les hypothèses diagnostiques soient envisagées, c'est-à-dire que la somme de leurs probabilités respectives soit de 100 %.
Dans une situation pour laquelle on connaît les probabilités *a priori* de différentes maladies M1, M2, M3, etc. la probabilité d'une maladie M1 si un signe S1 est présent s'écrit $p\ (M1|S1)$:

$$p(M1|S1) = \frac{p(M1) \times p(S1|M1)}{p(M1) \times p(S1|M1) + p(M2) \times p(S1|M2) + p(S1|M3) \text{ etc.}}$$

Il est encore possible d'utiliser les rapports de vraisemblance (RV) de chaque signe. Si la prévalence est de 50 %, la probabilité *a posteriori* (valeur prédictive positive) d'une maladie est égale au produit des RV apporté à ce produit + 1. Ainsi, une étude sur les gastro-entérites a

montré qu'elles étaient une fois sur deux d'origine virale. L'étude des RV en faveur de l'origine virale de l'affection a donné les résultats indiqués dans le tableau I [5].

Tableau I – Rapports de vraisemblance en faveur de l'origine virale d'une gastro-entérite (d'après [5]).

	Rapport de vraisemblance
Âge > 15 ans	1,8
Âge < 15 ans	0,6
Selles liquides	1,3
Selles molles	0,3
Vomissements	1,5
Absence de vomissement	0,4
Fièvre	0,9
Absence de fièvre	1,1
Rhinite	0,6
Absence de rhinite	1,3

Ainsi, un malade qui consulte pour une gastro-entérite et qui a moins de 15 ans, des selles liquides, des vomissements sans fièvre ni rhinite a une probabilité que sa diarrhée soit d'origine virale de :

$$\frac{1{,}8 \times 1{,}3 \times 1{,}5 \times 1{,}1 \times 1{,}3}{(1{,}8 \times 1{,}3 \times 1{,}5 \times 1{,}1 \times 1{,}3) + 1} = \frac{5{,}02}{5{,}02 + 1} = 0{,}83 \text{ ou } 83\ \%$$

En introduisant ainsi une à une, différentes variables, ce que l'on appelle l'utilisation séquentielle de plusieurs signes, et à condition que ces variables soient indépendantes entre elles, il est possible de concevoir des systèmes informatiques d'aide au diagnostic. Il ne s'agit là que d'une aide au diagnostic qui peut orienter la décision médicale, mais ne se substitue pas à elle. Les résultats obtenus, parfois décevants, avec des performances de l'ordinateur inférieures à celles du clinicien, peuvent être dus à des banques de données insuffisantes en nombre et en qualité. En revanche, de telles expériences ont montré que le fait d'obliger le clinicien à entrer des données précises d'interrogatoire et d'examen dans l'ordinateur l'amenait à interroger et à examiner les malades mieux et plus systématiquement qu'il ne le ferait autrement. Par exemple, dans l'aide au diagnostic sur les douleurs aiguës de l'abdomen, cela oblige à bien préciser les caractères d'une douleur : siège, irradiations, type, mode de début, etc. ou encore à ausculter l'abdomen afin de répondre aux interrogations correspondantes de

l'ordinateur. À ce titre, de tels systèmes ont une valeur pédagogique certaine. Ils peuvent aussi être utilisés de façon presque ludique en choisissant, parmi les différentes variables proposées, celle qui augmente le plus la probabilité d'un diagnostic donné et inversement. Autrement dit, il s'agit là d'une nouvelle façon d'apprendre la sémiologie.

Le choix d'un examen

Les examens complémentaires sont de plus en plus nombreux et sophistiqués. Les progrès techniques en diminuent les désagréments et les risques pour les patients. Aux angiographies qui nécessitaient jadis des injections intra-artérielles et la montée de cathéters se sont substituées des techniques de scannographie ou de résonance magnétique nucléaire qui ne demandent qu'une injection intraveineuse périphérique. La tentation est alors d'autant plus grande de prescrire facilement ces examens et de les utiliser de façon insuffisamment réfléchie. Les demandes étaient parfois motivées par le prétexte du « dossier complet », notion qui devait être démythifiée [1]. Aujourd'hui, ces examens sont souvent demandés par les médecins pour se couvrir dans la perspective d'une éventuelle plainte de malades dans l'esprit d'un principe de précaution. Néanmoins, la multiplication du nombre d'examens complémentaires que l'on peut demander est sujette à trois principales critiques.

Les contreparties de multiplier les examens complémentaires

La première contrepartie est d'ordre statistique. Comme il a été montré, à propos du risque de première espèce et des examens biologiques qui ont une distribution normale (cf. p. 98), plus on fait d'examens, plus on augmente le risque que l'un d'entre eux sorte des limites de la normale alors même que le sujet est normal. Rappelons que, si l'on prescrit six examens biologiques dont la distribution est normale et qui sont indépendants entre eux, il y a une chance sur quatre que l'un d'entre eux sorte des limites de la « normale ». Cela est vrai pour les examens biologiques comme pour les examens morphologiques avec des risques d'erreurs d'interprétation. Ainsi, il est loin d'être exceptionnel en pratique hospitalière de voir demander un examen radiologique, puis devant un résultat ambigu ou peu cohérent avec le contexte clinique, en demander un autre, en général de type différent (scanner après une échographie par exemple) puis, si le résultat du second n'est pas concordant avec celui du premier, un troisième examen, etc.

La seconde contrepartie d'augmenter le nombre d'examens complémentaires que l'on peut prescrire est, bien entendu, le coût. Même si cela paraît marginal par rapport à l'objectif de cet ouvrage, on ne peut pas ne pas rappeler à ce propos que leur prise en charge, en grande partie, par l'assurance maladie contribue à son déséquilibre avec une dette cumulée en 2006 qui s'élevait à 76 milliards et de 2009 à 2011, un déficit annuel d'un peu plus de 10 milliards.

La troisième critique que l'on peut formuler à la multiplication des examens complémentaires est qu'elle ne fait que refléter une absence ou une insuffisance de réflexion médicale et qui motive les aphorismes du tableau II. C'est, à nos yeux, la plus importante.

Tableau II – Aphorismes sur la multiplication des examens complémentaires.

En France, en l'absence de référence médicale opposable, tout médecin peut prescrire tous les examens biologiques, radiologiques, isotopiques possibles. Mais, seule la réflexion aboutit à faire des choix qui limitent la demande à une prescription « à bon escient ».
À l'hôpital, l'épaisseur et le poids d'un dossier médical ne sont pas toujours proportionnels à la réflexion des médecins.

Des demandes raisonnées d'examens complémentaires

La demande raisonnée d'examens complémentaires constitue certainement une composante difficile et, de ce fait, intéressante de l'exercice médical. En effet, la demande d'un examen complémentaire dans une démarche diagnostique doit s'intégrer dans une stratégie globale. Celle-ci doit prendre en compte la *valeur informationnelle* de l'examen, telle qu'elle a été définie dans les chapitres précédents, les résultats comparés des examens entre eux, leurs contreparties en termes de désagrégement ou même de risque pour le patient et de coût, le tout sans perdre de vue une perspective décisionnelle. « *Et quelle merveille quand l'intelligence, ainsi armée et entraînée, se met à fonctionner enfin comme un instinct !* [6] ».

La valeur informationnelle comparative des examens complémentaires entre eux

Si l'on dispose de plusieurs examens complémentaires pour une aide au diagnostic, ce qui est particulièrement le cas pour les examens

morphologiques, le choix de l'examen doit commencer par prendre en compte celui qui offre la meilleure valeur informationnelle, c'est-à-dire celui qui a la probabilité d'être le plus sensible, le plus spécifique, et d'avoir les meilleures valeurs prédictives. Ce choix repose sur les résultats comparant des examens entre eux dans le contexte clinique devant lequel on se trouve. Nous avons vu que les comparaisons qui apportaient le meilleur niveau de preuve de la supériorité d'un examen par rapport à un autre reposaient sur des essais randomisés ou mieux encore pour ces examens, sur leurs réalisations chez le même malade qui devient alors son propre témoin.

Mais le choix dépend encore de l'objectif médical qui est en jeu. S'il serait grave de passer à côté d'un diagnostic possible, il convient, en première intention, de demander un examen très sensible, même s'il est peu spécifique. Par exemple, dans une politique de dépistage des cancers colorectaux, il convenait de faire un examen comme l'Hémocult® qui dépiste l'existence de sang dans les selles avec une bonne sensibilité. Mais cet examen est assez peu spécifique, l'hémorragie pouvant être due à d'autres lésions du tube digestif. Il convenait alors, dans un second temps, de faire, dans la population sélectionnée par l'Hémocult®, un examen quasi spécifique qui est la coloscopie.

Un autre facteur qui doit être intégré dans la décision de prescription est le coût de l'examen. Un exemple est donné par une étude déjà ancienne, mais démonstrative. Elle avait comparé chez des malades qui avaient un cancer colorectal, l'échographie et la scannographie dans le diagnostic de métastases hépatiques [7]. Les résultats ont montré que l'échographie était un peu moins sensible que la scannographie, mais un peu plus spécifique, les différences n'étant pas statistiquement significatives (tableau III).

Tableau III – Valeurs comparées de l'échographie
et de la scannographie dans le diagnostic de métastases hépatiques
dans le cancer coloretal [7].

	Échographie	**Scannographie**	***p***
Sensibilité	82 %	91 %	ns
Spécificité	93 %	87 %	ns

Compte tenu de ces résultats, dans la mesure où le coût de l'échographie est environ cinq fois moindre que celui de la scannographie, une démarche diagnostique réfléchie en termes de coûts doit, chez les malades qui ont un cancer colorectal, limiter la recherche de métastases hépatiques à la prescription d'une échographie.

Mais la composante économique de la réflexion ne doit pas se limiter à l'intégration des seules deux données que sont la valeur informationnelle de l'examen et son coût unitaire. La connaissance de *données épidémiologiques* peut et doit encore intervenir. L'exemple de la radiographie pulmonaire avant une intervention chirurgicale est caricatural. Une radiographie thoracique coûtait environ 20 euros en 2010. Mais, chez un malade qui n'a pas de symptomatologie clinique cardiopulmonaire, des études ont montré que la probabilité de dépistage d'une anomalie par la radiographie thoracique était nulle chez les patients de moins de 20 ans et de trois sur cent après cet âge. Le coût du dépistage chez l'adulte se monte ainsi à environ 665 euros [8]. Ces mêmes études ont encore montré qu'une anomalie dépistée n'avait d'utilité décisionnelle que dans 0,2 % des cas. Le dépistage utile par une radiographie thoracique préopératoire revient ainsi à 333 000 euros. Le médecin est, bien entendu, libre de sa décision. Mais il doit intégrer ce raisonnement dans la prise de celle-ci. En ce qui concerne la radiographie thoracique avant une intervention, les travaux cités ont fait abandonner sa prescription systématique au profit d'un interrogatoire et d'un examen clinique qui ont été ainsi revalorisés [9].

L'importance de *cette utilité décisionnelle* peut encore être apportée par trois exemples. Le premier est banal. Si un adulte jeune a des douleurs abdominales et, à l'examen une contracture, les unes et l'autre prédominant dans la fosse iliaque droite, le diagnostic de péritonite appendiculaire est quasi certain (après s'être assuré, s'il s'agit d'une femme, qu'elle n'a pas d'antécédent ou de signes pouvant faire discuter une affection gynécologique). En principe, l'indication opératoire ne se discute pas. Il est alors inutile de chercher une hyperleucocytose. Qu'elle existe ou non (ce qui peut se voir dans une péritonite appendiculaire), la numération des leucocytes ne changera pas la décision.

Autre exemple, une étude prospective a été faite chez 24 malades qui avaient un cancer de la tête du pancréas afin d'estimer la valeur du Pet-Scan dans le diagnostic de métastases [10]. Les résultats ont montré que le Pet-Scan avait une sensibilité de 70 % et une spécificité de 83 %. Il était donc considéré par les biophysiciens qui avaient réalisé les examens comme un très bon examen. En fait, par rapport aux données de l'échographie, il n'avait modifié la décision que chez un seul malade. En admettant que le coût unitaire d'un Pet-Scan qui, à l'époque, était d'environ 1 500 euros, le coût d'un examen utile sur le plan décisionnel pouvait être estimé à 40 000 euros, ce qui a paru exorbitant.

Notre troisième exemple montre qu'il convient parfois de pousser assez loin le raisonnement et c'est bien ce qui fait l'intérêt de ce type de réflexion. Nous reprendrons l'exemple de l'échographie et de la scan-

nographie dans le diagnostic de métastases hépatiques qui semblait conclure à l'intérêt de l'échographie. En fait, si cette échographie ne montre pas de métastases hépatiques chez un malade qui a un cancer colique ou rectal, il doit être opéré afin de réséquer son cancer. La palpation du foie découvrira du reste, dans un faible pourcentage de cas, des petites métastases hépatiques passées inaperçues à l'échographie préopératoire dont la sensibilité n'est pas de 100 %. Si l'on découvre des métastases réséquables, elles pourront être réséquées dans le même temps que le cancer colique avec des probabilités de survie à cinq ans de l'ordre de 25 % [11]. S'il existe des métastases qui ne paraissent pas réséquables, il est encore souhaitable, de façon générale, d'opérer afin de réséquer le cancer colique primitif et de mettre le malade à l'abri de complications, hémorragies, occlusion intestinale, etc. Le diagnostic de métastases hépatiques avant l'intervention ne change donc pas, de façon générale, répétons-le, la décision chirurgicale. Dans la majorité des cas, il est ainsi possible de faire l'économie de leur recherche. De plus, la meilleure méthode de détection de ces métastases (standard de référence externe) est l'échographie pendant l'intervention avec biopsies des lésions suspectes.

Grille de réalisation ou de lecture des études sur l'évaluation d'un moyen diagnostique

Évaluation d'un moyen diagnostique

1. Les données fondamentales sont clairement précisées :

Population sur laquelle l'étude a porté.

Critères d'estimation du caractère normal ou pathologique du moyen diagnostique évalué.

Critères sur lesquels, dans la population étudiée, le diagnostic de maladie a été fait ou écarté.

2. Les effectifs des quatre sous-groupes qui en découlent sont donnés :

« Vrais positifs ».

« Faux positifs ».

« Faux négatifs ».

« Vrais négatifs ».

3. L'appréciation de la valeur du moyen diagnostique a été faite en termes de :

Sensibilité.

Spécificité.

Valeur prédictive positive.

Valeur prédictive négative.

(avec leurs intervalles de confiance).

4. Si le moyen diagnostique évalué est quantitatif :

Courbes ROC.

Comparaison de deux (ou plusieurs) moyens diagnostiques

1. Le sujet est son propre témoin ou, à défaut, essai randomisé.

2. Les risques de première et de seconde espèce ont été pris en compte.

3. Pour un moyen diagnostic quantitatif : test statistique de comparaison des courbes ROC.

Démarche diagnostique

Prise en compte de :

La valeur informationnelle de l'examen.

Les contreparties statistiques en termes de risque (notamment α et β).

Les contreparties médicales en termes de désagrément et morbidité.

Les contreparties économiques en termes de coûts.

In fine, l'utilité décisionnelle.

Il existe une liste d'items pour améliorer la qualité des publications sur les moyens diagnostiques (STARD, acronyme pour *Standard for reporting of diagnostic accuracy*).

Références

1. Mornex R (1977) Pour une stratégie des examens paracliniques. Nouv Presse Med 6: 1725-8
2. Palfrey S (2001) Daring to practice low-cost medicine in a high-tech era. New England J Med
3. AURC, ARC (1981) Les syndromes douloureux aigus de l'abdomen. Etude prospective multicentrique. Nouv Presse Med 10: 3771-3
4. AURC, ARC (1984) Aide au diagnostic et à la décision devant un syndrome douloureux abdominal aigu. Revue Epidémiol et Santé Pub 32: 40-4
5. Brachet R, Etienney I, Flahault A *et al.* (1999) Gastro-entérites hivernales. *Calicivirus* et *Rotavirus* ont été les deux familles de virus les plus fréquemment identifiées. Le Quotidien du médecin
6. De Romilly J (1998) Le trésor des savoirs oubliés. De Fallois, Paris, p 83
7. Alderson PO, Adams DF, McNeil BJ, *et al.* (1983) Computed tomography, ultrasound and scintigraphy of liver in patients with colon or breast carcinoma: a prospective comparison. Radiology 149: 225-30
8. Blery C (1980) Examens paracliniques pré-opératoires. Le Concours médical 102: 5607-10
9. National study by the Royal College of radiologists (1979) Preoperative chest radiology. Lancet 2: 83-6
10. Huguier M, Barrier A, Zacharias T, Valinas R (2006) Résultats de la tomographie par émission de positons dans les cancers de l'appareil digestif. Bull Acad Natle Med 190: 75-87
11. Weber JC, Bachellier P, Oussoulzoglou E, Jaeck D (2003) Simultaneous resection of colorectal primary tumour and synchronous metastases. Br J Surg 90: 956-62

Utilisation des méthodes multifactorielles dans une démarche diagnostique

4

Bien souvent la probabilité d'un diagnostic est liée à plusieurs données. Comme nous l'avons indiqué, elles peuvent être traitées en utilisant les probabilités bayésiennes qui, à partir d'une probabilité *a priori* reposant sur des données épidémiologiques, transforment, en introduisant une à une les différentes données dans le modèle, cette probabilité *a priori* en probabilité *a posteriori.*
Une autre méthode consiste à utiliser des méthodes multifactorielles prédictives et l'élaboration de scores.
Nous prendrons l'exemple de la lithiase vésiculaire et de la recherche d'une lithiase associée de la voie biliaire principale (VBP).
Les données du problème sont les suivantes :
– on sait que tout malade que a une lithiase vésiculaire peut avoir une lithiase associée de la VBP ;
– il est important de reconnaître ces lithiases de la VBP pour les traiter en même temps que la lithiase vésiculaire afin de mettre le malade à l'abri de complications comme un ictère, une angiocholite ou une pancréatite aiguë, mais les examens radiologiques habituels préopératoires sont très peu sensibles dans ce diagnostic :
  – aussi, le dépistage de lithiase de la VBP se faisait habituellement par une cholangiographie qui était le plus souvent réalisée au moment de l'intervention chirurgicale pour la lithiase vésiculaire. En France, le dogme était de faire systématiquement ces cholangiographies peropératoires lors de toute cholécystectomie pour lithiase vésiculaire. Or, on découvrait seulement chez 10 % des malades une telle lithiase de la VBP. On faisait donc inutilement une cholangiographie dans 90 % des cas.

Il pouvait alors paraître utile d'essayer de déterminer sur l'ensemble des malades qui avaient une lithiase vésiculaire l'existence d'un sous-groupe à très faible risque de lithiase de la VBP. Cela permettrait d'éviter de faire une cholangiographie peropératoire dans ce sous-groupe.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Une étude rétrospective a ainsi été faite chez 503 malades qui avaient eu une cholangiographie peropératoire pour dépister une lithiase de la VBP (variable expliquée) [1]. Onze covariables ont été étudiées. Huit d'entre elles étaient liées à l'existence d'une lithiase de la VBP en analyse unifactorielle. L'analyse multifactorielle, utilisant la régression logistique, a sélectionné cinq variables liées à la probabilité élevée d'une lithiase de la VBP : l'âge, des antécédents de colique hépatique, de cholécystite, la présence de calculs vésiculaires de moins de 10 mm, et une VBP de plus de 12 mm. Les résultats ont été exprimés par des *odds ratio* ajustés sur les autres variables. Par exemple, les patients ayant une VBP > 12 mm avaient 22 fois plus de risque d'avoir une lithiase de la VBP que ceux qui avaient une VBP < 12 mm et ce, indépendamment des autres covariables. Il a encore été possible de déterminer, à partir d'une équation de régression logistique (en utilisant le logarithme des *odds ratio*), un score dont l'équation est indiquée dans le tableau I.

Tableau I – Score de probabilité, chez un malade qui a une lithiase vésiculaire, de lithiase associée de la voie biliaire principale.

0,03 × âge (en années).
+ 2,2 si la voie biliaire principale a > 2 mm de large.
+ 1,5 si les calculs vésiculaires ont < 10 mm.
+ 0,7 si le malade a des antécédents de colique hépatique.
+ 0,8 s'il a une cholécystite.
Si une covariable est absente, on la cote 0.

À partir de ces données, un score pouvait être calculé pour chaque malade. Puis, il a été possible de déterminer les valeurs du score pour lesquelles la probabilité d'une lithiase de la VBP était très faible, moyenne ou élevée (tableau II).

Tableau II – Probabilité d'une lithiase de la voie biliaire principale en fonction de la valeur du score prédictif calculé.

Score	Lithiase de la voie biliaire principale		
	Oui effectif	Non effectif	Valeur prédictive positive (%)
≥ 5,9	50	12	81
< 5,9 > 3,5	35	169	17
≤ 3,5	5	232	2

Les résultats de cette étude ont ainsi suggéré que, chez les malades dont le score était inférieur à 3,5, la probabilité de lithiase de la VBP était de

2 %. Dans ce sous-groupe, la pratique systématique d'une cholangiographie peropératoire amènerait à faire 98 examens inutiles pour deux examens utiles. Il a donc été décidé de ne plus faire de cholangiographie dans ces cas qui représentaient dans cet échantillon, près de la moitié des malades (237 sur 503).
Dans les études de ce type, il est toujours souhaitable, sinon nécessaire, de tester le modèle proposé sur d'autres échantillons de malades que ceux qui ont servi à élaborer le score. C'est ce qui a été fait dans notre exemple dans deux autres études qui ont validé ce modèle [2, 3].

Références

1. Huguier M, Bornet P, Charpak Y, *et al.* (1992) Selective contraindications based on multivariate analysis for operative cholangiography in biliary lithiasis. Surg Gynecol Obstet 172: 470-4
2. Montariol T, Rey C, Charlier A, *et al.* (1995) Preoperative evaluation of the probability of common bile duct stones. French Association for Surgical research. J Am Coll Surg 172: 470-4
3. Millat B, Deleuze A, de Saxce B, *et al.* (1997) Routine intraoperative cholangiography is feasible and efficient during laparoscopic cholecystectomy. Hepato-gastroenterol 44: 22-7

Concordance

5

Ce chapitre sur la concordance aurait pu être placé dans la seconde partie de cet ouvrage qui concerne les comparaisons. Néanmoins, la mesure de concordance étant surtout utilisée en médecine pour comparer l'interprétation par deux praticiens d'une même série d'examens, par exemple radiologiques ou histologiques, il nous a paru plus logique d'en expliquer le principe dans ce chapitre.
La concordance s'applique, bien entendu, à d'autres problèmes comme celui d'apprécier si les notes données par deux correcteurs à des mêmes copies d'examen (ou les notes d'un même correcteur à deux moments différents) sont cohérentes entre elles. La concordance peut encore être utilisée pour apprécier la valeur de deux examens différents, par exemple l'examen cytologique et l'examen d'anatomopathologie de la moelle osseuse dans la recherche de cellules cancéreuses circulantes. Si la concordance entre les deux examens est bonne, cela permet de choisir celui qui est le plus simple à réaliser ou le moins onéreux.
De façon générale, la concordance a pour but d'apprécier s'il y a similitude ou non entre deux ou plusieurs informations se rapportant au même objet. Elle apporte une information différente et complémentaire de celles données par la sensibilité, la spécificité ou les valeurs prédictives d'un examen.

Ce que n'est pas la concordance

La concordance diffère d'une relation statistique

Les tests statistiques permettent de savoir si une différence entre deux résultats est significative ou non ou encore dans les études d'équivalence, s'il n'y a pas de différence ; dans les deux cas par le rejet de l'hypothèse nulle H0. Autrement dit, ils apprécient l'association

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

qui peut exister entre différentes variables. Mais ils ne permettent pas d'apprécier une concordance.

En voici une illustration.

Supposons que deux cliniciens examinent, de façon indépendante l'un de l'autre, 100 malades et fassent les diagnostics suivants d'appendicite aiguë (tableau I).

Tableau I – Résultats (fictifs) du diagnostic de deux examinateurs chez 100 malades qui ont une suspicion d'appendicite aiguë.

	Docteur Galien :	
	Pas d'appendicite	***Appendicite***
Docteur Vésale :		
Pas d'appendicite	50	0
Appendicite	0	50

Le calcul du χ^2 montre une valeur de 100, ce qui correspond à une valeur de $p < 0{,}001$. Il y a donc une relation statistiquement significative entre les diagnostics des deux examinateurs. Il y a aussi une parfaite concordance entre eux (nous verrons que κ est égal à 1).

Supposons, dans un second exemple tout aussi fictif, que les résultats d'une autre étude concernant 100 autres malades examinés par deux autres examinateurs, toujours de façon indépendante l'un de l'autre soient les suivants (tableau II).

Tableau II – Résultats (fictifs) du diagnostic de deux examinateurs chez 100 malades qui ont une suspicion d'appendicite aiguë.

	Docteur Hippocrate :	
	Pas d'appendicite	***Appendicite***
Docteur Hunter :		
Pas d'appendicite	0	50
Appendicite	50	0

Dans ce second exemple, le χ^2 est de 100 ($p < 0{,}001$). Il existe une relation statistiquement significative entre les diagnostics des deux examinateurs. En revanche, il y a une discordance totale entre les deux examinateurs (nous verrons que κ est égal à – 1).

La concordance n'est pas l'appréciation d'une proportion identique d'événements.

Prenons, cette fois l'exemple, toujours fictif, de deux radiologues qui voient une série d'échographies et dont les conclusions sont les suivantes (tableau III) :

Tableau III – Résultats (fictifs) du diagnostic de deux radiologues à l'examen de N échographies.

	Docteur Roentgen :	
	Examen normal	***Examen anormal***
Docteur Écho :		
Examen normal	*a*	*b*
Examen anormal	*c*	*d*

La comparaison des pourcentages de diagnostic des deux radiologues peut être évaluée à l'aide d'un test statistique χ^2 apparié de McNemar, comme le montre le tableau IV.

Tableau IV – χ^2 apparié de McNemar.

$$\frac{(|\text{Discordance pour Roentgen} - \text{discordance pour Écho}| - 1)^2}{\text{Nombre total de discordances}},$$

ce qui donne : $\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$

Comme pour le χ^2 non apparié, si ce χ^2 est supérieur à 3,84, la différence est statistiquement significative.

La comparaison de ces résultats et le test statistique ne permettent pas d'évaluer s'il y a ou non concordance entre les deux radiologues. En effet, les malades reconnus à juste raison par le docteur Roentgen ne sont pas nécessairement les mêmes que ceux reconnus par le docteur Écho.

La concordance

La concordance brute

La concordance brute ou pourcentage d'agréments répond à une notion simple : c'est la proportion observée de diagnostics identiques chez deux examinateurs. Dans l'exemple du tableau I, elle est de 100 % ; dans celui du tableau II, elle est de 0 %. Dans l'exemple du tableau III, elle est égale à :

$$(a + d) / (a + b + c + d).$$

Cette mesure n'est pas très sensible aux divergences qui peuvent exister si les effectifs sont très déséquilibrés entre les classes utilisées dans la cotation.

Le coefficient kappa (κ) [1]

Le coefficient kappa permet, comme les tests d'inférence statistique, de faire la part du hasard dans les résultats d'une étude sur la concordance. Prenons l'exemple de l'interprétation de 106 scannographies par deux radiologues (tableau V).

Tableau V – Interprétation de 106 scannographies par deux radiologues. Effectifs observés.

	Interprétation du docteur White :		
	Examen normal	***Examen anormal***	***Total***
Docteur Black :			
Examen normal	56	12	68
Examen anormal	8	30	38
Total	64	42	106

La concordance brute observée (*Po*) est la somme des résultats concordants : 56 + 30 = 86 rapportée à l'effectif total : 106, soit 81 %.
En fait, cette concordance brute est la résultante de la concordance réelle et de la concordance liée au hasard de l'échantillon. L'index κ permet en quelque sorte « d'expurger » de la concordance brute la part du hasard. Notons qu'il existe des extensions au kappa pour plus de deux observateurs et quel que soit le nombre de classes étudiées.
Pour le docteur White, le pourcentage de résultats normaux est de 64/106 soit 60 %. Si l'on reporte ce pourcentage à l'ensemble des examens scannographiques trouvés normaux (n = 68) par le docteur Black, l'effectif des résultats normaux pour celui-ci serait 60 % de 68, soit 41. Il est possible de calculer de façon analogue les effectifs des trois autres cases du tableau, ou plus simplement de soustraire 41 des effectifs des lignes et des colonnes, ce que montre le tableau VI.

Tableau VI – Interprétation de 106 scannographies par deux radiologues. Effectifs calculés.

	Interprétation du docteur White :		
	Examen normal	***Examen anormal***	***Total***
Docteur Black :			
Examen normal	41	27	68
Examen anormal	23	15	38
Total	64	42	106

La concordance attendue par hasard (*Pa*) est égale à 41 + 15/106, soit 53 %.
Cela peut s'écrire de façon plus générale (tableau VII).

Tableau VII – Effectifs calculés.

	Examinateur A :		
	Examen normal	***Examen anormal***	***Total***
Examinateur B :			
Examen normal	a'	b'	$l1$
Examen anormal	c'	d'	$l2$
Total	$c1$	$c2$	N
l1, l2, c1, c2, ayant les mêmes valeurs que les valeurs observées. $a' = (l1 \times c1) / N$. $d' = (l2 \times c2) / N$.			

Le coefficient κ estime le rapport entre, au numérateur, la concordance observée et la concordance calculée ou attendue (*Po – Pa*) et au dénominateur, le complément de la concordance attendue (1 – *Pa*) (tableau VIII).

Tableau VIII – Le calcul du coefficient kappa (κ).

$$\kappa = \frac{Po - Pa}{1 - Pa}$$

ce qui donne dans notre exemple :

$$\kappa = \frac{0{,}81 - 0{,}53}{1 - 0{,}53} = \frac{0{,}28}{0{,}47} = 0{,}60$$

Les valeurs de κ peuvent être comprises entre – 1 et + 1. Le tableau IX indique les qualificatifs qui correspondent usuellement à différentes valeurs de κ (échelle de Landis et Koch [2]).

Tableau IX – Qualificatifs usuels en fonction de la valeur de κ (d'après [2]).

Valeurs de kappa	**Concordance considérée comme :**
1	parfaite
0,81 à 0,99	excellente
0,61 à 0,80	bonne
0,41 à 0,60	modérée
0,21 à 0,40	faible
0,00 à 0,20	très faible
< 0,00	désaccord

Dans notre exemple de l'examen scannographique, le coefficient kappa de 0,60 peut faire considérer que la concordance est modérée, alors que la concordance brute observée était de 0,81 et pouvait la faire considérer comme excellente.
L'utilité du coefficient kappa peut être montrée à partir de deux autres exemples dans lesquels la concordance brute observée est similaire alors que les coefficients kappa diffèrent.

Tableau X – Premier exemple.

	Interprétation			
	de l'examinateur A		**Effectifs calculés**	
	Examen		**Examen**	
	anormal	***normal***	***anormal***	***normal***
Interprétation de l'examinateur B :				
Examen anormal	80	10	81	9
Examen normal	10	0	9	1

Dans ce premier exemple, la concordance observée est de 80 %. La concordance calculée est de 82 %, proche de la concordance observée. Le coefficient kappa est égal à 0,11, c'est-à-dire que la concordance est presque nulle.

Tableau XI – Second exemple.

	Interprétation			
	de l'examinateur A		**Effectifs calculés**	
	Examen		**Examen**	
	anormal	***normal***	***anormal***	***normal***
Interprétation de l'examinateur B :				
Examen anormal	40	10	25	25
Examen normal	10	40	25	25

Dans ce second exemple, la concordance observée est de 80 %, bien qu'elle reflète une situation très différente de la précédente. La concordance calculée est seulement de 50 %. Le coefficient kappa est égal à 0,60, c'est-à-dire que la concordance est modérée.

Le kappa pondéré

Dans l'évaluation de la concordance, des mesures effectuées par deux examinateurs avec une échelle qui comporte plus de deux niveaux ordonnés, on peut juger que tous les désaccords n'ont pas le même poids. Prenons l'exemple de deux anatomopathologistes qui évaluent le degré de fibrose dans le foie avec le score METAVIR, avec un jugement en trois catégories de fibrose allant de F2 (fibrose modérée) à F4 (cirrhose). On peut considérer qu'un écart d'une unité, par exemple sur une même lame, une cotation de F3 par un lecteur et de F4 par l'autre, est moins grave qu'un écart de deux unités.

Supposons que ces deux anatomopathologistes aient examiné les mêmes malades et aient abouti aux conclusions suivantes (tableau XII).

Tableau XII – Résultats observés par deux anatomopathologistes sur une même série de lames.

	Anatomopathologiste A			
	Lésion			
	F2	**F3**	**F4**	**Total**
Anatomopathologiste B				
F2	20	5	0	25
F3	4	25	6	35
F4	2	10	28	40
Total	26	40	34	100

Le calcul du kappa donnerait comme résultat $\kappa = (20 + 25 + 28)/100 - (34,1/100)/(1 - 34,1/100) = 0,59$. Cette approche considère que seule l'identité parfaite entre les cotations marque l'accord entre les lecteurs. Plus le nombre de catégories augmente, plus il sera donc difficile de parvenir à une bonne concordance. On peut considérer que cette évaluation est trop drastique et préférer choisir de pondérer les erreurs de telle sorte qu'un désaccord d'un point soit considéré comme presque concordant et moins grave qu'un désaccord de 2 points, par exemple.

Le kappa pondéré permet cette approche : on va tenir compte de la concordance et dans les lectures où des désaccords existent, on leur donne un poids d'autant plus faible que l'écart est fort. Dans notre exemple, parce que la cotation comporte trois niveaux effectivement utilisés, on donnera par exemple un poids de 1 en cas d'accord parfait, un poids de ½ en cas de différence d'un point et un poids de 0 en cas de différence de deux points. Plus généralement on donnera un poids inverse à la distance entre deux cotations, s'étalant de 1 en cas d'accord parfait à 0 en cas de désaccord le plus grand.

Le calcul du kappa pondéré est détaillé dans le tableau XIII.

Tableau XIII – Calcul du kappa pondéré à partir de l'exemple du tableau XII.

1) Pourcentage de concordance observé :

Avec accord parfait $\frac{20 + 25 + 28}{100} = 73\ \%$

Avec désaccord de 1 unité $\frac{4 + 10 + 5 + 6}{100} = 25\ \%$

Avec désaccord de 2 unités $\frac{2 + 0}{100} = 2\ \%$

2) Effectifs attendus :

	Anatomopathologiste A			
	Lésion			
	F2	F3	F4	Total
Anatomopathologiste B				
F2	6,5	10	8,5	25
F3	9,1	14	11,9	35
F4	10,4	16	13,6	40
Total	26	40	34	100

3) Pourcentage de concordance dû au hasard :

Avec accord parfait $\frac{6{,}5 + 14 + 13{,}6}{100} = 34{,}1\ \%$

Avec désaccord de 1 unité $\frac{9{,}1 + 16 + 10 + 11{,}9}{100} = 47\ \%$

Avec désaccord de 2 unités $\frac{10{,}4 + 8{,}5}{100} = 18{,}9\ \%$

4) Le kappa pondéré est alors égal à :

$$\frac{(0{,}73 + 0{,}5 * 0{,}25 + 0 * 0{,}02) - (0{,}341 + 0{,}5 * 0{,}47 + 0 * 0{,}189)}{1 - (0{,}341 + 0{,}5 * 0{,}47 + 0 * 0{,}189)} = 34{,}1\ \%$$

ce qui donne :

$$\frac{0{,}855 - 0{,}576}{1 - 0{,}576} = 0{,}66.$$

Comme attendu, la valeur du kappa pondéré est plus élevée, puisqu'il donne de l'importance à toutes les réponses, et pas uniquement celles qui sont identiques entre lecteurs.

Conclusions

L'analyse de concordance permet de déterminer, parmi plusieurs paramètres histologiques d'une maladie, ceux qui prêtent le moins à des difficultés d'interprétation, c'est-à-dire ceux dont le coefficient kappa est le plus élevé [4].

Références

1. Cohen J (1960) A coefficient of agreement for nominal scales : Educational and psychological measurement. 20: 37-46
2. Landis JR, Koch GG (1977) The measurement of observer agreement for categorial data. Biometrics 33: 159-74
3. Henk JM, Kunkler PB, Smith CW (1977) Radiotherapy and hyperbaric oxygen in head and neck cancer. Lancet 2: 101-3
4. Chastang C, Césarini YP, Beltzer-Garelli H, *et al.* (1984) Établissement d'une classification pronostique en deux stades du mélanome malin primitif à partir d'une analyse multidimensionnelle et d'une étude de concordance. Rev Epidém et Santé Publi 32: 243-8

Partie 5

Les évaluations thérapeutiques

Introduction

Certains traitements, par leur grande efficacité et leurs effets indésirables limités, ont fait rapidement l'objet d'un consensus. Ce fut le cas du traitement du diabète par l'insuline, de la leucémie de l'enfant par la chimiothérapie, de la maladie ulcéreuse duodénale par les inhibiteurs de la pompe à protons ou de l'angor par les dérivés nitrés.

Mais les progrès thérapeutiques sont souvent moins évidents. L'empirisme thérapeutique qui se fonde sur des résultats observés sur un groupe de malades, sans groupe témoin, souvent rétrospectivement, résultats que l'on peut qualifier d'anecdotiques, est trop critiquable sur le plan scientifique pour continuer à être accepté aveuglément. L'histoire de la thérapeutique est ainsi jonchée de traitements que l'on a cru être efficaces et qui, pour cette raison, ont été largement prescrits mais qui, en définitive, ne se sont pas avérés plus efficaces qu'un placebo. Ainsi, le pourcentage de traitements considérés comme « efficaces » après des études non ou mal contrôlées, n'utilisant pas d'insu, avant de s'avérer ultérieurement inefficaces ou mêmes nocifs, a été estimé à près de 50 % [1].

Bien souvent, le médecin, dans ses prescriptions thérapeutiques est amené à faire un choix entre plusieurs médicaments ou encore entre une attitude médicale et une intervention chirurgicale ou bien encore entre deux techniques chirurgicales.

Nous avons évoqué les comparaisons rétrospectives ou historiques cherchant à guider ces choix en évaluant, par exemple, un nouveau traitement administré à une série de patients par rapport à un autre traitement plus ancien ou de référence pour lequel on dispose de données recueillies ou publiées antérieurement. Ces comparaisons doivent être interprétées avec la plus grande prudence et ne permettent guère de conclusions, les chances étant trop faibles pour que les traitements que l'on cherche à comparer aient été administrés à des groupes de malades similaires. Dans ces études comparatives, il existe en effet des biais qui entraînent presque inéluctablement des différences dans les

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

résultats, dues à d'autres facteurs que les traitements que l'on cherche à comparer. Ainsi, dans les comparaisons dites « historiques », comparant un groupe de malades anciens et de malades traités de façon plus récente, des biais sont, par exemple, liés au fait que les malades les plus récents ont des affections dont le pronostic spontané est meilleur grâce aux progrès du dépistage ou encore que des traitement(s) associé(s) à celui que l'on cherche à évaluer sont devenus plus efficaces.
Dans les comparaisons qui portent sur deux groupes de malades, vus pendant la même période, le biais habituel est lié au fait qu'il existe des raisons qui ont généralement motivé le fait que certains malades aient reçu un traitement et les autres un traitement différent.
Dans ces deux types de comparaisons, on peut donc être quasi certain que la comparaison des traitements a porté sur des groupes de malades qui n'étaient pas similaires.
Nous avons vu que seuls les essais randomisés, menés selon les bonnes pratiques, garantissaient l'absence de ces biais. De ce fait, ils apportent le meilleur niveau de preuve dans la prise de décision.
Nous ne reviendrons pas sur ce qui a été dit dans la deuxième partie concernant les comparaisons et particulièrement les essais randomisés. Ces essais sont encore appelés essais contrôlés parce qu'il y a un groupe contrôle qui est comparé au groupe recevant le nouveau traitement que l'on cherche à évaluer.

Référence

1. Venning GR (1982) Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarm. Br Med J 284: 249-52

Les comparaisons thérapeutiques ne reposant pas sur des essais randomisés

1

Le but de ce chapitre est de montrer tous les aléas, les erreurs de jugement que nous venons d'évoquer et que les comparaisons thérapeutiques qui ne reposent pas sur les résultats d'essais randomisés risquent d'entraîner et de faire commettre.

Les études non contrôlées

Le médecin avait et a encore l'habitude de fonder sa décision sur l'enseignement qu'il a reçu, ses lectures ultérieures, son expérience, sa formation permanente et sur les messages qui lui sont apportés par les représentants des firmes pharmaceutiques. Il prescrit en conséquence le traitement qui lui paraît être le plus efficace, le mieux toléré, et éventuellement – pas assez souvent –, le moins onéreux. De ce dernier point de vue, l'absence de prescription médicamenteuse à la suite d'une consultation médicale, fréquente dans certains pays industrialisés de très bon niveau médical, s'oppose à la pratique française qui fait de la France le plus gros consommateur de médicaments au monde après les États-Unis.

Ces comportements sont fondés sur la conviction du médecin, ce qui est normal. Mais elle n'est pas toujours, loin s'en faut, fondée sur le meilleur niveau de preuve scientifique existante. Même si la conviction du médecin n'est pas dénuée d'intérêt, elle diffère d'une évaluation scientifique.

En voici un exemple. Le traitement de l'ulcère gastro-duodénal par des complexes ferrico-ferro-sodiques a longtemps été considéré comme relativement efficace puisqu'il s'associait dans près de 50 % des cas à une régression de la poussée ulcéreuse. Mais avec la pratique des endoscopies gastro-duodénales dans des essais randomisés comparant les anti-H2 et un placebo, on s'est aperçu qu'environ 50 % des malades qui recevaient le placebo guérissaient spontanément. Le taux

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

de guérison de près de 50 % avec les complexes ferrico-ferro-sodiques correspondait donc au taux spontané de guérison de ces ulcères et leur prescription a été abandonnée au profit des anti-H2, puis des inhibiteurs de la pompe à proton, à la suite d'essais thérapeutiques randomisés.
Un autre exemple concerne la duodéno-pancréatectomie céphalique, résection qui est habituellement réalisée dans les cancers de la tête du pancréas. Dans cette intervention, après l'exérèse, le chirurgien doit effectuer une anastomose entre le pancréas corporéo-caudal restant et le tube digestif. Il peut utiliser pour cela l'estomac ou le jéjunum. Le principal risque est celui d'une fistule de cette anastomose. On a pensé et espéré qu'un traitement préventif par la somatostatine, en inhibant la sécrétion pancréatique serait susceptible de réduire ce risque de fistule. De nombreux chirurgiens ont donc prescrit à leurs opérés de la somatostatine dont le coût journalier était élevé, de plusieurs centaines d'euros. Or, en 1997, un essai randomisé ayant porté sur 120 opérés a comparé un groupe traité par de la somatostatine à un groupe témoin [1]. Il a montré que le taux de fistules pancréatiques n'était pas statistiquement différent dans les deux groupes. Cet exemple montre que la logique qui découle de connaissances biologiques, biochimiques, microbiologiques, etc. inspire certains progrès thérapeutiques, mais ne suffit pas à les prouver.

Les comparaisons « historiques »

Les comparaisons « historiques », bien que d'un niveau de preuve très inférieur à celui des essais thérapeutiques randomisés, peuvent cependant suffire à elles-mêmes en cas de progrès thérapeutiques très importants. Elles ont ainsi fait la preuve de leur utilité comme ce fut le cas du traitement des leucémies myéloïdes chroniques par le Glivec®. Avant, ces leucémies se transformaient toujours en leucémies aiguës rebelles à tout traitement, ce qui est devenu exceptionnel. Ces comparaisons historiques ont ainsi permis de faire bénéficier tous les malades de ce traitement beaucoup plus rapidement que si l'on avait dû attendre les résultats d'un essai randomisé. Malheureusement, bien des progrès thérapeutiques sont moins évidents ou plus hypothétiques. Les comparaisons des résultats d'un traitement « ancien » avec ceux obtenus par un traitement plus récent sont cependant sujettes à des erreurs d'interprétation. Deux exemples le montrent clairement.
Le premier concerne la cholécystectomie par cœlioscopie. Les avantages attribués à cette technique par rapport à la cholécystectomie par mini-laparotomie étaient une durée d'hospitalisation plus brève et une reprise plus rapide de l'activité professionnelle. Or, un essai randomisé

comparant ces deux techniques a été réalisé [2]. Les médecins qui prenaient la décision de la sortie du malade de l'hôpital et de la reprise de ses activités professionnelles ne savaient pas le type d'intervention qui avait été faite, cœlioscopie ou mini-laparotomie. L'étude n'a pas montré de différence entre les deux traitements. Il est donc probable que la durée d'hospitalisation plus brève et la reprise d'activité plus rapide après chirurgie par cœlioscopie n'étaient pas liées à la technique, mais à la conviction des chirurgiens qu'*a priori* la cœlioscopie permettait de faire sortir plus rapidement les opérés de l'hôpital et de leur faire reprendre plus tôt une activité normale. Cette étude montre, par ailleurs, la supériorité des études randomisées en simple ou en double insu sur les accords professionnels et les conférences de consensus, même si la contrainte expérimentale est parfois difficile à respecter.

Un autre exemple concerne la duodéno-pancréatectomie céphalique dont nous avons déjà parlé. Des données biologiques et expérimentales avaient suggéré qu'après exérèse, l'anastomose du pancréas corporéo-caudal restant avec l'estomac exposerait moins au risque de fistule que l'anastomose avec le jéjunum. De fait, une comparaison historique sur une soixantaine de malades n'a montré aucune fistule après anastomose pancréatico-gastrique alors qu'il en avait été observé 17 % après anastomose pancréatico-jéjunale [3]. Un essai randomisé a néanmoins été réalisé comparant ces deux techniques [4]. Il a montré que le taux de fistules était similaire dans les deux groupes infirmant les résultats de la comparaison rétrospective. En fait, en lisant les résultats de la première étude portant sur une comparaison historique, on s'apercevait que, dans les premières années de cette étude, il avait été surtout fait des anastomoses pancréatico-jéjunales alors que dans les dernières années, il avait été surtout fait des anastomoses pancréatico-gastriques. Les meilleurs résultats obtenus avec cette dernière technique s'expliquaient donc probablement par une expérience plus importante des chirurgiens et non pas par le type de dérivation pancréatique.

Dans les comparaisons historiques, même si les auteurs cherchent à vérifier *a posteriori* que les deux groupes qui ont été comparés sont bien similaires, des différences peuvent toujours passer inaperçues. Il en est ainsi d'évolutions thérapeutique marginales par rapport au critère de jugement principal.

Études prospectives non randomisées

Ces études ont deux objectifs principaux. Le premier est la faisabilité d'un traitement, notamment en cancérologie. Elles permettent ainsi de rejeter rapidement, sans grand risque de se tromper, un traitement dont l'efficacité paraît trop faible pour être évaluée par un essai

randomisé, toujours difficile, long et coûteux à réaliser. Les études prospectives non randomisées permettent également, dans des essais en cancérologie ou dans des maladies rares, d'analyser la pharmacocinétique des nouvelles molécules (doses, métabolisme, etc.) ; ce sont les essais dits de phase 2.
Le principal intérêt de ces études est le recueil prospectif des données qui, par rapport au recueil rétrospectif, réduit les risques d'avoir des données manquantes et de perdre de vue certains malades, ce qui complique encore l'interprétation des résultats.
Ces études sont encore la meilleure façon, en vue d'un essai randomisé, de se faire une opinion objective de ce que l'on peut attendre d'un nouveau traitement par rapport à un traitement de référence pour estimer le nombre de sujets qu'il sera nécessaire d'inclure dans l'essai randomisé afin de limiter le risque de deuxième espèce (cf. p. 92).
Néanmoins, lorsque ces études prospectives sont comparatives, mais non randomisées, il n'y a pas de garantie que les deux sous-groupes chez lesquels on cherche à comparer deux traitements seront similaires. En effet, des raisons ont presque toujours motivé le fait que certains sujets ont reçu un traitement et les autres, un autre traitement. Ces comparaisons sont ainsi très critiquables, sauf exception.

L'effet placebo

Dans les essais randomisés, si l'on cherche à apprécier l'efficacité d'un traitement, il est souhaitable de constituer un groupe témoin qui reçoit un placebo du traitement actif, le placebo désignant une substance pharmacologiquement inerte. Ce placebo doit avoir le même aspect physique que le médicament. La nécessité de comparer l'action du médicament avec celle d'un placebo, évoquée à propos du traitement de l'ulcère duodénal se justifie par l'existence, qui a été démontrée, d'un effet placebo. Par exemple, sur quinze études ayant concerné un peu plus de mille patients ayant des phénomènes douloureux, un placebo d'antalgique était efficace en moyenne dans 35 % des cas [5]. À titre anecdotique, cette étude a paradoxalement montré que les douleurs organiques angoissantes étaient celles qui répondaient le mieux au placebo.
L'effet placebo peut cependant avoir un substratum biologique comme cela a été montré par des études expérimentales chez l'animal. Il en est de même chez l'homme. Par exemple, dans la maladie de Parkinson, un essai randomisé comparant la L-DOPA à un placebo a montré au Pet-Scan que l'effet placebo observé correspondait à une libération de dopamine dans le striatum [6].

Dans tout essai thérapeutique randomisé, le substratum organique de l'effet placebo justifie d'autant plus la comparaison médicament-placebo. C'est sur ce type de comparaison que se fonde la Commission de transparence qui dépend de la Haute autorité de santé pour évaluer le bénéfice apporté par des nouveaux médicaments. En fait, il se dégage une tendance de cette Commission à demander des comparaisons entre un médicament de référence et un nouveau médicament plutôt qu'entre un nouveau médicament et un placebo.

L'amélioration des études observationnelles

Les études observationnelles sont une première étape, souvent indispensable, à toute évaluation. Encore faut-il qu'elles soient de bonne qualité méthodologique. Tel a été l'objectif de la grille de réalisation de ces études (*Strengthening the reporting of observational studies in epidemiology* ou STROBE) [7].

Grille d'évaluation méthodologique d'une étude observationnelle [7]

Cette grille s'inspire de : *Strengthening the reporting of observational studies in epidemiology,* STROBE.

Nous avons mis **en gras**, ce qui nous paraît à la fois particulièrement important et souvent en défaut.

1. Exposé des données qui ont motivé l'étude.

2. **Les données fondamentales :**

a. Sujets inclus dans l'étude :

– critères d'inclusion et d'exclusion ;

– nombre de sujets remplissant les critères d'inclusion, mais non entrés dans l'essai et raisons ;

– description de l'échantillon ;

– étude rétrospective ou prospective ?

b. Ce que l'on cherche à évaluer :

– appareil d'investigation, dispositif médical implantable etc. (fabricant, date) ;

– ou traitement médical (posologie, mode et horaires d'administration, autres traitements admis ou non) ;

– ou traitement chirurgical (technique) ;

– en cas d'événement indésirable, ce qui est prévu ?

3. **Les critères de jugement :**

– principal ;

– secondaires ;

– recueil par qui et comment (en insu) ?

4. Analyse des résultats.

– **déviations par rapport au protocole** (inclus secondairement exclus, allocation de protocole erronée etc.) ; jugement en intention de traiter, puis *per* protocole.

– perdus de vue ;

– analyse de sous-groupes ;

– intervalles de confiance à 95 %.

5. Considérations éthiques et réglementaires.

– consentement éclairé ;

– promotion et obligations légales.

6. Lors de l'élaboration du protocole :

– date de début et de fin espérée des inclusions ;

– financement.

Références

1. Lowy AM, Lee JE, Pisters PW, *et al.* (1997) Prospective, randomized trial of octeotride to prevent pancreatic fistula after pancreaticoduodenectomy for malignant disease. Ann Surg 226: 632-41
2. Majeed AW, Troy G, Nicholl JP, *et al.* (1996) Randomized, prospective, single-blind comparison of laparoscopic versus small-incision cholecystectomy. Lancet 347: 989-94
3. Mason GR, Freeark RJ (1995) Current experience with pancreatogastrostomy. Am J Surg 169: 217-9
4. Yeo CJ, Cameron JL, Maher MM, et al.(1995) A prospective randomized trial of pancreaticogastrostomy versus pancreaticojejunostomy after pancreaticoduodenectomy. Ann Surg 222: 580-92
5. Beecher HK (1955) The powerful placebo. JAMA 1602-6
6. De La Fuente-Fernandez R, Ruth TJ, Sossi V, *et al.* (2001) Expectation and dopamine release: mechanism of the placebo effect in Parkinson's disease. Science 293: 1164-6
7. Von Elm E, Altman DG, Egger M, *et al.* (2007) The strengthening the reporting of observational studies in epidemiology (STROBE) Statement. Guideliness for reporting observational studes. PloS Med 4

Lorsqu'un essai randomisé n'est pas possible 2

Si les essais thérapeutiques randomisés sont le moyen qui offre le plus de garanties scientifiques pour fonder son opinion sur un choix thérapeutique, des raisons techniques ou éthiques peuvent limiter leur réalisation [1].
Les comparaisons historiques ou l'étude dite « ici, ailleurs » permettent de constituer un contrôle. Cependant, il faut s'assurer de la comparabilité initiale des groupes d'intervention et de ces groupes contrôle. Les études multifactorielles, parce qu'elles permettent un ajustement sur les différences entre groupes, permettent une meilleure comparabilité entre l'intervention et le contrôle. L'ajustement peut, notamment, être réalisé par la technique du score de propension où l'on comparera les résultats chez des sujets qui ont reçu des interventions différentes alors même qu'ils avaient la même probabilité initiale de recevoir l'une ou l'autre.

Les études multifactorielles

Voici un exemple dans lequel un essai randomisé n'était pas envisageable et pour lequel une étude multifactorielle a permis d'apporter une assez bonne réponse à la question posée. Il s'agissait du traitement chirurgical du cancer du tiers moyen du rectum qui avait été défini précisément dans sa hauteur sur le rectum [2]. À l'époque de cette étude, deux traitements étaient pratiqués. Le traitement de référence était l'amputation abdomino-périnéale du rectum et de l'appareil sphinctérien qui impliquait la constitution d'une colostomie terminale définitive dans la fosse iliaque gauche. Peu à peu, une meilleure connaissance de l'extension, en général très limitée vers le bas, de ces cancers du tiers moyen du rectum et des progrès techniques ont permis de les réséquer en faisant une anastomose colo-anale qui a l'avantage d'éviter aux malades une colostomie. Néanmoins, on pouvait craindre que ces exérèses moins étendues, notamment vers le bas, par rapport

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

aux amputations, augmentent le risque de récidives et de décès. Des comparaisons sur des données rétrospectives comportaient trop de biais pour permettre d'en tirer des enseignements peu contestables.

Théoriquement, un essai randomisé aurait été souhaitable pour savoir ce qu'il en était et si la résection-anastomose ne diminuait pas les chances de survie des malades. Mais la clause d'ambivalence impliquait de n'inclure que des malades qui pouvaient avoir, dans de bonnes conditions carcinologiques et techniques, soit une amputation, soit une résection-anastomose. Il aurait alors été impossible sur le plan éthique, à la suite d'un tirage au sort, de faire à un malade une amputation avec colostomie définitive, alors que celle-ci pouvait être techniquement évitée.

La solution a été de faire une étude multifactorielle. Le critère de jugement principal a été la survie. Le modèle de Cox a été utilisé en incluant, parmi les covariables, le type d'intervention réalisée et en regardant si, à hauteur constante des autres covariables, ce type d'intervention réalisé était ou non lié à la survie.

Sur 119 malades inclus dans l'étude, il a été d'abord fait une étude unifactorielle portant sur les covariables qui pouvaient être liées à la survie, incluant le type d'intervention réalisée. Ensuite, en ne retenant, parmi les covariables que celles qui avaient une incidence statistiquement significative sur la survie (logrank), une analyse multifactorielle a été faite en utilisant le modèle de Cox. Le risque relatif de la résection-anastomose par rapport à l'amputation a enfin été estimé en réalisant un ajustement sur les covariables qui étaient liées à la survie en analyse multifactorielle. Il était de 1,05 pour la survie et de 0,78 pour les récidives locales. Ces risques relatifs n'étaient pas statistiquement significatifs avec une bonne puissance des tests.

Les conclusions que l'on pouvait tirer d'une telle étude étaient les suivantes :

1. Bien entendu, le niveau de preuve obtenu par cette étude n'atteignait pas celui qu'aurait eu un essai randomisé.
2. L'absence de différence statistiquement significative entre les deux modalités thérapeutiques ne démontrait pas pour autant l'équivalence comme cela a été expliqué dans le chapitre concernant les comparaisons et il aurait mieux valu faire un essai d'équivalence qu'un essai randomisé « classique ».
3. Il était néanmoins possible d'estimer que, s'il y avait différence entre les deux traitements, elle était de faible ampleur.

Des travaux ont visé à améliorer la qualité des publications d'études non randomisées. Le plus diffusé est probablement le *Transparent reporting and evaluating with non-randomized designs* (TREND)[1].

1 www.trend-statement.org:asp/trend/asp.

Il comporte une liste de 22 items, mais, contrairement à d'autres instruments d'évaluation, il ne comprend pas de scores.

Les scores de propension

Cette méthode est couramment utilisée en économétrie. Le score de propension désigne la probabilité, pour une personne de caractéristiques données, d'être exposée à un traitement. La distribution de ce score sur les groupes de traitements comparés fournit un critère de jugement de la comparabilité entre ces deux groupes. S'il y a un biais de recrutement, les scores auront tendance à être élevés pour les patients exposés et faibles pour les non exposés. Afin de neutraliser ce biais au maximum, un sous-échantillon de patients comparables entre les deux groupes peut être élaboré, par appariement sur les scores de propension : chaque patient exposé au « nouveau traitement » est apparié au patient du groupe témoin ayant le score le plus proche, à condition que la différence entre les deux scores ne soit pas trop grande. Ce sous-échantillon possède ainsi des caractéristiques proches de l'essai clinique. Toutefois, il ne permet d'assurer une similitude des groupes que sur les caractéristiques observées.

Un exemple récent concerne la mortalité après gastroplastie de réduction chez les obèses. Un objectif de ces interventions est de réduire la mortalité de ces personnes, à distance de l'intervention, en leur évitant le développement de pathologies liées à l'obésité. Dans un travail, les auteurs ont comparé 850 patients ayant subi une gastroplastie à une population contrôle de 41 255 personnes [3]. La mortalité observée dans le groupe d'obèses opérés, six ans après l'intervention était de 7 % contre 15 % dans le groupe contrôle. Mais les opérés étaient plus souvent plus jeunes et de sexe féminin que les sujets du groupe contrôle. Les auteurs ont alors calculé un « score de propension », c'est-à-dire la probabilité d'avoir subi une intervention selon l'âge, le sexe, le poids, etc. À l'issue de ce calcul, il leur a été possible de sélectionner 847 opérés et 847 témoins ayant le même score de propension. Ces deux groupes étaient alors similaires en tout point, sauf la gastroplastie. La comparaison de la mortalité entre ces deux groupes a montré une mortalité de 7 % dans chaque groupe suggérant que la gastroplastie ne semblait pas réduire la mortalité par comorbidité chez les obèses.

Signalons qu'il existe d'autres nouvelles approches de modélisation statistiques comme les modèles « causal » utilisant les « graphes dirigés acycliques » ou les « modèles structuraux marginaux ».

La différence entre les études multifactorielles et les scores de propension est que les premières incluent dans l'analyse finale des résultats

tous les malades de l'étude, alors que les scores de propension n'incluent qu'une partie du groupe témoin, sélectionnée par des caractéristiques aussi proches que possible de celles du groupe traité. Les études multifactorielles sont plus applicables quand il n'y pas de différence quantitativement importante entre les deux groupes, et les scores de propension quand le groupe témoin est beaucoup plus important que le groupe d'intérêt.

La recherche d'un consensus : la méthode « Delphi »

En l'absence de résultats d'essais randomisés ou d'analyses multifactorielles, un pis-aller est la méthode Delphi. Son nom vient de la ville de Delphes, en Grèce, où la pythie, oracle d'Apollon, faisait ses prédictions. Le principe de la méthode repose sur les réponses d'un groupe d'experts à des séries de questionnaires préétablis. Après chaque série de questions, une synthèse des réponses est remise à chacun d'entre eux avec les arguments sur lesquels sont fondées ces réponses. Ensuite, il est demandé aux experts de revoir leurs réponses à la lumière de la synthèse qu'ils ont reçue. Ce processus est répété, souvent trois ou quatre fois. Les réponses sont anonymes pour limiter le risque que les idées de certains experts, dont l'aura est importante, influencent trop et l'emportent sur celles des autres. En général, on constate qu'à la suite de ce processus, les divergences s'estompent et convergent vers un certain consensus.

La qualité finale du résultat repose, en définitive, sur celle des experts et sur la capacité des analystes dans le traitement des réponses et de la conduite de tout l'exercice.

Le processus peut également déboucher sur le constat qu'un consensus est impossible dans l'état actuel des connaissances, et pointer ainsi les études à mener en priorité.

Par exemple, une enquête a porté sur la prescription des radiographies en réanimation dans une trentaine de cas cliniques [4]. Cette enquête a montré que les praticiens étaient, notamment, en désaccord sur le recours à la radiographie systématique et quotidienne chez les patients sous ventilation mécanique. Pour cette raison, un essai randomisé a été réalisé. Il a montré que la radiographie à la demande n'entraînait pas de perte de chance pour le patient, tout en réduisant son irradiation et les coûts [5].

Références

1. Solomon MJ, McLeod RS (1995) Should we be performing more randomized controlled trials evaluating surgical operations? Surgery 118: 459-67
2. Huguier M, Chastang C, Houry S, *et al.* (1997) Sphincter-saving resection or not for cancer of the midrectum. Am J Surg 174: 11-5
3. Maciejewski ML, Livingston EH, Smith VA (2011) Survival among high-risk patients after bariatric surgery. JAMA 305(23): 2419-26
4. Hejblum G, Loos V, Vibert JF, *et al.* (2008) A web-based Delphi study on the indications of chest radiography for patients in ICUs. Chest 133: 1107-12
5. Hejblum G, Chalumeau-Lemoine L, Loos V, *et al.* (2009) Comparison of routine and on-demand prescription of chest radiography in mechanically ventilated adult: a multicentre, cluster-randomized, two-period cross-over study. Lancet 374: 1687-93

Revue systématique et méta-analyses des essais randomisés

3

Les méta-analyses constituent une revue systématique de la littérature scientifique, dans laquelle l'accent est mis sur une synthèse quantitative des résultats. Le matériau de base de la méta-analyse est l'étude randomisée recherchée de façon exhaustive et systématique [1] ; c'est-à-dire qu'elles doivent prendre en compte tous les essais randomisés ayant inclus des malades aux caractéristiques similaires et ayant comparé des traitements, eux aussi similaires avec les mêmes critères de jugement. De même que pour un essai randomisé classique, l'objectif est d'estimer l'efficacité ou non d'un traitement par rapport à un autre et l'importance des différences observées là où les études existantes, prises une à une, donnent des résultats divergents [1]. Elles diffèrent d'une simple sommation de ces études par l'introduction d'un facteur de pondération. De plus, en rassemblant la totalité des données disponibles, les méta-analyses permettent, en principe, de déceler des effets indésirables rares qui n'auraient pas été vus dans chaque étude.

La revue systématique ou méta-analyse doit répondre à des critères de qualité similaires à ceux des essais randomisés. Au lieu d'inclure des malades comme dans un essai randomisé, on y inclut des essais thérapeutiques. On doit y retrouver les données fondamentales des essais randomisés : critère d'inclusion et de non-inclusion des études que l'on analyse, définition des traitements qui sont comparés, définition des critères de jugement.

Ces exigences de qualité des revues systématiques reposent au départ sur la similitude des essais inclus, ce qui est loin d'être évident.

1. Les populations incluses dans les essais doivent être similaires. Par exemple, dans des études de chimiothérapie dans les cancers colorectaux réséqués, des malades chez lesquels l'examen anatomopathologie de la pièce d'exérèse a montré des métastases ganglionnaires, mais qui n'ont pas de métastases viscérales apparentes.

1 Des logiciels comme Revman® ont pour but la recherche d'articles en vue d'une méta-analyse.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7,

2. Les traitements qui ont été évalués dans chaque essai doivent être les mêmes ou peu différents les uns des autres afin que l'interprétation de leur efficacité ne soit pas entachée d'ambiguïté. Dans notre exemple de chimiothérapie adjuvante, le ou les produits utilisés, les posologies, les modes d'administration doivent être les mêmes ou très peu différents et s'il y a des différences d'un essai à l'autre qui paraissent acceptables pour que l'essai soit quand même inclus dans la méta-analyse, bien entendu, ces différences doivent, être signalées.
3. Il faut enfin que les essais aient les mêmes critères de jugement, du moins le même critère de jugement principal. Il faut encore qu'ils aient été estimés de la même façon ce qui est évident pour une survie, mais ne l'est pas toujours pour une récidive ou un dosage biologique.

Toutes ces conditions sont rarement réunies. Néanmoins, il est alors possible de faire des méta-analyses plus globales afin de dégager des tendances générales, par exemple en incluant tous les essais randomisés portant sur une même classe thérapeutique et non pas sur un même traitement. Plus rarement, à défaut de comparaison directe entre deux traitements A et B, il est possible de comparer les résultats de deux méta-analyses, l'une comparant le traitement A à un placebo, l'autre le traitement B à un placebo. L'utilisation des *odds ratio* permet alors, à défaut de comparaison directe entre le traitement A et le traitement B, de se faire une opinion sur la supériorité éventuelle d'un traitement par rapport à l'autre.

Les biais rencontrés dans les méta-analyses

Une méta-analyse sur un sujet ne doit pas inclure les essais randomisés qui ont été mal conduits, ni, comme nous venons de le voir, les essais sur le sujet qui ont été réalisés avec des critères d'inclusions ou de jugement différents. Les raisons d'exclusion de ces essais de la méta-analyse doivent alors être indiquées.

Le biais le plus important des méta-analyses est éditorial : la sélection de celles-ci par les périodiques médicaux. En effet, si l'on peut raisonnablement admettre que tous les essais randomisés dont les résultats sont positifs ont fait l'objet d'une publication, en revanche il faut toujours craindre que des essais de recherche de supériorité dont les résultats sont négatifs ne soient pas publiés ou seulement sous forme de résumés de communication à des congrès, résumés qui ne sont pas indexés sur les sites de banques informatiques de données. Ce biais de non-publication amène à surestimer l'effet d'un traitement. Ainsi, une étude a comparé les résultats d'une méta-analyse qui avait inclus 21 essais randomisés qui avaient été publiés avec ceux de 29 études sur

le même sujet qui avaient été enregistrées, mais dont huit n'avaient pas été publiées [2]. La méta-analyse des essais publiés montrait une différence statistiquement significative en faveur d'un traitement (p = 0,02), alors que la méta-analyse de l'ensemble des 29 essais ne montrait pas de différence en faveur de ce traitement (p = 0,25).

Un autre biais est représenté par les publications multiples d'un même essai randomisé. Bien que cette pratique ne soit pas éthiquement acceptable, des auteurs peu scrupuleux peuvent s'y livrer. Par exemple, une recherche sur des essais randomisés qui ont été publiés concernant l'ondanstéron a montré que c'était le cas pour 17 % d'entre eux [3]. Ces publications redondantes sont parfois difficiles à dépister lorsqu'elles sont réalisées dans des langues différentes avec des auteurs mis dans des ordres différents, voire qui ne sont pas toujours les mêmes.

C'est la raison pour laquelle, l'opinion qui consiste à estimer que les méta-analyses constituent le niveau de preuve le plus élevé de la médecine factuelle est sujette à caution à cause de ces biais qui ne sont pas toujours détectables, loin s'en faut.

Les registres d'essais randomisés qui existent, par exemple, en France, à l'Agence française de sécurité sanitaire des produits de santé ont pour principal objectif d'informer la communauté scientifique des essais en cours, d'éviter des essais redondants et de limiter un risque de biais par défaut des méta-analyses.

Hétérogénéité des essais randomisés inclus dans une méta-analyse

Si l'inclusion de tous les essais randomisés dans une méta-analyse est indispensable, un déséquilibre risque de se produire si les effectifs de malades inclus dans les différents essais sont très différents les uns des autres. Les études qui ont inclus le plus grand nombre de malades « tirent vers elles » l'ensemble des résultats. La méta-analyse s'éloigne alors d'autant plus de ce qui était son objectif.

De plus, une méta-analyse n'a de sens que si les résultats des différents essais analysés ne sont pas diamétralement opposés c'est-à-dire s'il n'y a pas d'hétérogénéité qualitative. Dans le cas contraire, la méta-analyse est un instrument inadapté. La réflexion doit alors porter sur la compréhension des causes qui sont susceptibles d'expliquer de telles divergences de résultats, ce qui peut être une source intéressante d'enseignements.

Néanmoins, le plus souvent, il n'existe que des différences limitées dans l'efficacité d'un traitement d'un essai à l'autre. C'est l'hétérogénéité quantitative que l'on ne peut évoquer que si les différences d'un essai à

l'autre varient plus que ce que les fluctuations d'échantillonnage pouvaient laisser prévoir. Des tests, peu puissants, peuvent mesurer cette hétérogénéité. Ils permettent de vérifier l'absence de différence significative entre les résultats de chaque essai par rapport aux autres. En pratique, l'étude de la représentation graphique des *odds ratio* permet une évaluation suffisante de l'homogénéité relative des résultats des différents essais inclus dans la méta-analyse.

En l'absence d'hétérogénéité, une méthode (de Peto) permet de calculer les rapports de cotes selon un modèle à effet fixe.

L'évaluation des résultats : l'utilisation des *odds ratio*

L'utilisation des *odds ratio* trouve une excellente application dans les méta-analyses.

Prenons l'exemple fictif d'une méta-analyse ayant porté sur trois essais thérapeutiques randomisés (tableau I).

Tableau I – Exemple fictif d'une méta-analyse portant sur trois essais randomisés.

Essais	Groupe traité		Groupe contrôle	
	Nombre de malades		Nombre de malades	
	vivants	décédés	vivants	décédés
A	37	2	41	1
B	151	11	146	4
C	102	5	101	7

À partir de ces données, il est possible de calculer pour chaque essai l'*odd ratio* associé aux décès et son intervalle de confiance. Le tableau II montre ce calcul pour le premier essai (A).

Tableau II – Calcul de l'odds ratio du premier essai (A).

L'*odds* des décès du groupe traité est égal à 2/37 (pt) = 0,05.

L'*odds* des décès du groupe contrôle traité est égal à 1/41 (pc) = 0,02.

L'*odds ratio* est égal à : $\frac{2/37}{1/41} = \frac{2 \times 41}{37 \times 1} = 2{,}2$.

Le tableau III montre le calcul de la variance de cet *odds ratio*, l'écart-type, et l'intervalle de confiance.

Tableau III – Variance de l'écart-type et de l'intervalle de confiance de l'*odds ratio*.

$$\text{Variance } (s^2) = \frac{pt\,(1 - pt)}{nt} + \frac{pc\,(1 - pc)}{nc} = \frac{0{,}05\,(1 - 0{,}05)}{39} + \frac{0{,}02\,(1 - 0{,}02)}{42} = 0{,}0016$$

L'écart type est égal à $\sqrt{0{,}0016} = 0{,}041$.

Intervalle de confiance à 95 % de l'*odds ratio* = *odds ratio* ± (1,96 × 0,041), soit 2,2 ± 0,08

ce qui s'écrit :

	odds ratio	Intervalle de confiance à 95 %
Essai A	2,2	2,192 – 2,280

De la même façon, on peut calculer les *odds ratio* pour l'essai B et l'essai C avec leurs intervalles de confiance.

Souvent ces résultats sont exprimés sous forme de graphiques sur lesquels l'axe des abscisses représente la valeur de l'*odds ratio* de chaque essai et de l'analyse combinée de tous les essais. On nomme ce graphique « *forest plot* ». Le trait vertical représente la valeur 1 de l'*odds ratio*. La valeur estimée de chaque *odds ratio* est indiquée par un carré ou un losange dont la taille est proportionnelle aux effectifs de l'essai. Enfin, les traits horizontaux expriment l'intervalle de confiance de l'*odds ratio* (fig. 1).

Le tableau IV montre un exemple de méta-analyse sur la prévention des fistules digestives par l'enrobage de l'anastomose avec une colle biologique [4].

Tableau IV – Méta-analyse sur la prévention des fistules digestives après résection-anastomose par une colle biologique.

	Fistules		**Odds ratio**	**Intervalle de confiance 95 %**
	Groupe témoin	**Groupe traité**		
Étude : 1	1/28	1/32	1,15	0,07-18,88
2	6/49	6/57	1,16	0,36-3,93
3	8/94	3/51	1,45	0,40-5,23
4	5/52	2/48	1,47	0,71-3,06
Total	20/223	12/188	1,47	0,71-3,06

Cet exemple montre qu'il y a moins de fistules dans le groupe traité que dans le groupe témoin. Ce résultat suggère une homogénéité relative. Néanmoins, aussi bien pour chaque essai que dans le résultat global de la méta-analyse, les différences ne sont pas statistiquement

significatives comme le prouve bien le fait que tous les intervalles de confiance englobent la valeur 1.
La figure 1 montre la traduction graphique du tableau IV.

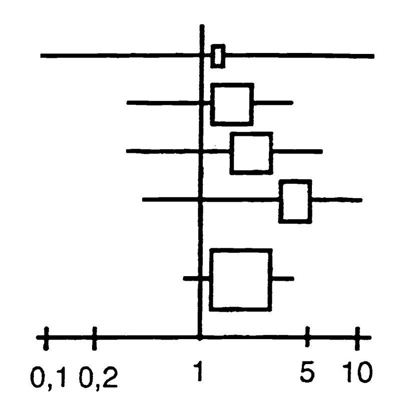


Fig. 1 – Représentation graphique de la valeur des *odds ratio* de chaque essai randomisé inclus dans la méta-analyse et de leur intervalle de confiance. Dans cet exemple, l'échelle des abscisses qui représente la valeur des *odds ratio* est logarithmique.

Qualité des méta-analyses

Une liste d'items a été élaborée sur les critères de qualité des méta-analyses [5] (*Quality of reporting of meta-analyses*, QUORUM). Un diagramme permet notamment d'analyser le nombre d'essais retrouvés pour la méta-analyse ainsi que le nombre et les raisons d'exclusion de certains essais.

Références

1. Pignon JP, Poynard T (1991) Méta-analyse des essais thérapeutiques. Gastroenterol Clin Biol 15: 229-238
2. Simes J (1986) Publication bias: the case of an international registry of clinical trials. J Clin Oncol 4: 1429-1441
3. Tramèr MR, Reynolds DJM, Moore RA, McQuay HJ (1997) Impact of covert duplication on meta-analysis: a case study. Br Med J 315: 635-640
4. Urbach DR, Kennedy ED, Cohen HM (1999) Colon and rectum anastomosis do not require routine drainage. A systematic review and meta-analysis. Ann Surg 229: 174-180
5. Moher D, Cook JD, Eastwood S, *et al.* (1999) Improving the quality of reports of meta-analyses of randomised controlled trials : the QUORUM statement. Lancet 354: 1896-900

Choix d'un traitement

4

Le choix d'un traitement doit intégrer :
- le bénéfice et les contreparties médicales que l'on peut attendre d'un traitement par rapport à ceux d'autres traitements, voire à l'abstention thérapeutique dans certains cas ;
- le coût direct du traitement et celui éventuel de ses effets adverses. Ces analyses de coût-avantage relèvent aujourd'hui de la médecine.

Il convient de signaler dans le cadre de la médecine factuelle l'aide que peut constituer pour répondre à une question de pratique clinique le *Patient intervention control outcome* (PICO)[1]. Le P concerne le patient et le problème qu'il pose ; I l'intervention évaluée ; C la comparaison par rapport à un autre traitement ; O le critère de jugement des résultats.

Bénéfices et contreparties médicales des traitements

Les comparaisons qui portent sur des traitements entre eux doivent faire la balance entre les avantages attendus et les contreparties médicales lorsque celles-ci ne sont pas négligeables. Nous avons donné l'exemple d'un essai randomisé sur la chimiothérapie intra-artérielle des métastases hépatiques des cancers colorectaux comparée à la chimiothérapie intraveineuse [1]. Cet essai randomisé a montré un bénéfice statistiquement significatif sur la survie (logrank). Mais le gain de médiane de survie n'était que de quatre mois et, en contrepartie, dans le groupe de chimiothérapie intra-artérielle, il y a été observé près de 60 % d'hépatites chimiques ou de cholangites sclérosantes dues à la chimiothérapie. Pour cette raison, la voie d'administration intra-artérielle, avec les antimitotiques alors utilisés a été abandonnée. Mais il est habituel que le rapport bénéfice/contreparties médicales soit nettement en faveur du traitement le plus efficace.

1 http//askmedline.nlm.nith/ask/pico.php.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Dans ces évaluations, il est souvent souhaitable d'inclure, parmi les critères de jugement, la qualité de vie qui prend en compte la perception par le patient de son propre état de santé grâce à des instruments de mesure qui ont été validés. Parmi ces instruments, on distingue des instruments génériques[2] et des instruments spécifiques, adaptés à la maladie et à son traitement. Par exemple, dans les maladies chroniques inflammatoires de l'intestin, c'est le cas de l'*Inflammatory Bowel Disease Questionnaire*. Il en est d'autres qui sont adaptés à un groupe de pathologies comme le *Gastrointestinal Quality of Life Index* pour les maladies digestives ou le *Quality of Life Questionnaire Core 30 Items*, décliné par organe en cancérologie[3].

Les études de coût-avantage

Ces études se décomposent en études coût-bénéfice, coût-efficacité, coût-utilité

Les études coût-bénéfice sont destinées à relier les coûts aux conséquences exprimées en unités monétaires. Elles consistent à évaluer, par rapport à une situation de référence (Ref), la somme des différences entre les coûts des prises en charge dans cette situation de référence (Cref) et les coûts en cas de traitement (Tt) (CTt). Ces différences de coûts sont évaluées pour chaque dépense d'hospitalisation (C1), de médecine de ville (C2), médicamenteuse (C3), etc. On calcule la différence entre le coût du Tt et celui de la situation de Ref (tableau I).

Tableau I – Études coût-bénéfice.

Différences de coûts de prise en charge : (C1Tt – C1Ref) + (C2Tt – C2Ref), etc. où : – C1Tt est le coût d'un type de frais (par exemple hospitalisation) chez les sujets traités ; – C1Ref est le coût de ce même type de frais chez les sujets de référence ; – C2Tt est le coût d'un autre type de frais (par exemple soins extrahospitaliers) chez les sujets traités ; – C2Ref est le coût de cet autre type de frais chez les sujets de référence.
Différence liée au coût du traitement : CTt – Cref, c'est-à-dire le coût du traitement moins le coût chez les sujets de référence (qui est nul s'il n'a pas été traité).
Agrégation coût/bénéfice absolue = (C1Tt – C1Ref) + (C2Tt – C2Ref), etc. – (CTt – Cref).

2 Comme le *Medical Outcomes Study (MOS) 36 items Short Form* ou SF-36.

3 www.quolid.org du *MAPI Research Institute*.

Ceci peut être illustré par un exemple[4]. Il concerne, chez la femme ménopausée, l'évaluation, par rapport à l'absence de traitement, de deux stratégies de traitement hormonal, l'une pendant quinze ans, l'autre à vie, afin d'essayer de diminuer le risque d'ostéoporose et de fractures du col du fémur. Les coûts comparés ont concerné les frais d'hospitalisation pour traitement d'une fracture du col du fémur, les frais de soins à domicile, ceux en institution et les frais du traitement hormonal. Les données sont indiquées dans le tableau II.

Tableau II – Coûts de prévention chez 100 000 femmes ménopausées d'une fracture du col du fémur par un traitement hormonal, en millions d'euros. Les données.

	Coûts			
	Hospitalisation	**À domicile**	**En institution**	**Traitement hormonal**
Pas de traitement hormonal	167	156	2 236	0
Traitement pendant 15 ans	142	141	2 186	129
Traitement à vie	75	121	2 138	282

À partir de ces données, on peut calculer les coûts et les bénéfices des deux stratégies de traitement par rapport à l'absence de traitement (tableau III).

Tableau III – Coûts et bénéfices des deux stratégies thérapeutiques par rapport à l'absence de traitement.

	15 ans par rapport à rien
Bénéfice	(167 – 142) + (156 – 141) + (2 236 – 2 186) = 90
Coût	129 – 0 = 129
	Traitement à vie par rapport à rien
Bénéfice	(167 – 75) + (156 – 121) + (2 236 – 2 138) = 225
Coût	282 – 0 = 282

L'agrégation coût-bénéfice absolue est égale à la différence entre le coût et le bénéfice (tableau IV).

4 D'après Van der Loos, Thèse de doctorat, Lausanne, 1984.

Tableau IV – Agrégation coût-bénéfice absolue.

Stratégie 15 ans de traitement : 90 – 129 = – 39
Stratégie de traitement à vie : 225 – 282 = – 57

Cet exemple montre que l'agrégation coût-bénéfice absolue de la stratégie du traitement sur 15 ans est préférable à celle du traitement à vie. Il est encore possible d'évaluer l'agrégation coût-bénéfice relative qui est égale au rapport coût/bénéfice (tableau V).

Tableau V – Agrégation coût-bénéfice relative.

Stratégie 15 ans de traitement : 129/90 = 1,43
Stratégie de traitement à vie : 282/225 = 1,25

Cette fois, c'est la stratégie du traitement à vie qui et préférable.
Ainsi, la stratégie agrégation coût-bénéfice absolue de traitement sur 15 ans est celle qui procure le gain le plus élevé, c'est-à-dire la plus grande différence entre le bénéfice et le coût. La stratégie coût-bénéfice relative de traitement à vie est celle pour laquelle le quotient coût sur bénéfice est plus faible.
La divergence entre les résultats de ces deux types d'estimation suggère que c'est alors la discussion qu'elle suscite qui peut être intéressante.

Les études coût-efficacité

Ces études ont pour objectif de relier les coûts d'un traitement à ses résultats en termes de santé et exprimés en unités physiques comme le nombre d'année de vies sauvées, le nombre de maladies évitées, etc. Si une stratégie est la plus efficace en termes de santé et de gain, elle s'impose comme le bon choix. Si deux stratégies ont la même efficacité, l'étude coût-efficacité permet de choisir celle qui offre le gain le plus élevé au nom de la minimisation des coûts.
Mais il y a parfois divergence. Dans notre exemple précédent, il a été montré que les espérances de vie à 50 ans étaient les suivants (tableau VI) :

Tableau VI – Espérance de vie à 50 ans (en jours).

	Espérance de vie	Gains d'efficacité
Pas de traitement hormonal	12 143	–
Traitement pendant 15 ans	12 163	20
Traitement à vie	12 206	63

La stratégie « traitement pendant 15 ans » apporte un gain d’efficacité de 20 jours pour un coût de 39 millions pour 100 000 femmes (tableau IV). L’autre stratégie « traitement à vie » apporte un gain d’efficacité de 63 jours pour un coût de 57 millions. Il est donc logique de choisir cette deuxième stratégie en termes d’efficacité. Mais c’est celle qui coûte le plus cher.

Il existe alors deux méthodes d’aide à la décision : l’agrégation coût-efficacité en moyenne et l’agrégation coût-efficacité marginale. Leur objectif est, en se ramenant à une efficacité similaire des deux stratégies, de comparer leurs coûts respectifs. Pour ce faire, il convient de supposer que, pour chaque traitement, le gain ou le coût est fonction du niveau d’efficacité atteint.

L’agrégation coût-efficacité en moyenne suppose que cette fonction soit de type linéaire. Il est donc possible de calculer le gain d’une unité d’efficacité et le traitement choisi sera celui dont le gain d’une unité d’efficacité sera le plus important.

En reprenant notre exemple, le gain d’une unité d’efficacité pour le traitement de 15 ans est de : 39/20 = 1,95. Le gain d’une unité d’efficacité pour le traitement à vie est de : 57/63 = 0,90. L’agrégation coût-efficacité amène ainsi à préférer la stratégie du traitement de 15 ans.

En fait, l’hypothèse de linéarité est très rarement satisfaite sur le plan économique, d’une part du fait de l’existence de coûts fixes, d’autre part de l’impossibilité d’accroître indéfiniment l’efficacité d’une stratégie. Ainsi, dans les cas fréquents où l’hypothèse de linéarité ne peut être satisfaite, il est possible d’utiliser l’agrégation coût-efficacité marginale.

L’agrégation coût-efficacité marginale suppose que les gains et les efficacités de deux traitement que l’on compare soient proches l’un de l’autre. Par ailleurs, il est nécessaire de décider pour lequel des deux niveaux d’efficacité la comparaison des gains sera effectuée. Dans notre exemple, le gain d’efficacité du traitement à vie est supérieur à celui du traitement pendant 15 ans. Si l’on compare les deux traitements en partant du niveau d’efficacité du traitement pendant 15 ans, la procédure coût-efficacité marginale consiste à estimer quelle serait la dépense générée par une unité supplémentaire d’efficacité pour le traitement de 15 ans, puis la dépense pour atteindre le niveau d’efficacité du traitement à vie.

Par exemple si la stratégie « traitement pendant 15 ans » génère un coût de 1 000 et une efficacité de 100 et si la stratégie « traitement à vie » génère un coût de 1 200 pour une efficacité de 105, pour gagner une unité d’efficacité avec la première stratégie, il faudrait dépenser 1 000/100 = 100 et pour atteindre 105, il faudrait une dépense de 1 500. À efficacité similaire, c’est alors la stratégie « traitement à vie » qui sera privilégiée.

Les études coût-utilité relient les coûts d'une action médicale à des critères qui ne sont plus monétaires, mais à ses conséquences exprimées en termes médicaux, par exemple les équivalents d'années de vie gagnées, pondérées par la qualité. On peut ainsi comparer dans les cancers épidermoïdes du canal anal, la chirurgie et la radiothérapie ou dans le reflux gastro-œsophagien, le traitement médical par les inhibiteurs de la pompe à proton et la chirurgie.

Référence

1. Rougier P, Laplanche A, Huguier M, *et al.* (1992) Hepatic arterial infusion of floxuridine in patients with liver metastases from colorectal carcinoma: long-term results of a prospective randomized trial. J Clin Oncol 10: 1112-8

Partie 6

Les évaluations pronostiques

Introduction

La connaissance d'un pronostic permet de prédire la probabilité d'une évolution chez un malade, par exemple 75 % de chances de survie à cinq ans après une intervention chirurgicale pour un cancer. Cette notion est probabiliste : chez un malade déterminé, si la connaissance de ce pronostic permet d'estimer qu'il a plus de chances d'être en vie au bout de cinq ans que l'inverse, il n'est pas possible pour autant de savoir s'il sera parmi les 75 % de survivants ou bien parmi les 25 % de malades décédés.

L'intérêt de la connaissance d'un pronostic ne se limite pas à ce type de prédiction individuelle. Cette connaissance permet encore d'adapter un traitement en tenant compte du pronostic. Ainsi, dans les cancers du côlon, après une exérèse chirurgicale apparemment complète, le taux de survie à cinq ans, en l'absence de métastases ganglionnaires à l'examen anatomopathologique, est de l'ordre de 80 %. Mais il est inférieur s'il existe des métastases ganglionnaires. Dans ces cas, les essais thérapeutiques randomisés ont montré qu'une chimiothérapie augmentait la survie. Ainsi, la connaissance de la valeur pronostique des métastases ganglionnaires et les résultats des essais randomisés font prescrire une chimiothérapie chez les malades qui ont été opérés d'un cancer du côlon et qui ont des métastases ganglionnaires. Cependant, ces chimiothérapies ont des contreparties notamment digestives et hématologiques. Pour cette raison, deux orientations dans les recherches caractérisent la logique des démarches médicales dans ce domaine. Certaines recherches se font vers des chimiothérapies, au moins aussi efficaces sur des essais d'équivalence, mais mieux tolérées et moins astreignantes. D'autres recherches cherchent à identifier, parmi les malades qui ont des métastases ganglionnaires, un sous-groupe à faible risque de récidive chez lequel on pourrait éviter, pour cette raison une chimiothérapie. C'est le cas d'études génomiques ou des micro-satellites.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Dans le même état d'esprit, un troisième intérêt des études pronostiques est de contribuer aux évaluations thérapeutiques par les essais randomisés. En effet, la connaissance d'un facteur de pronostic déterminant doit amener à réaliser une stratification sur ce facteur pronostique.
Enfin, la connaissance des facteurs pronostiques permet de réduire le nombre d'examens complémentaires qui peuvent être réalisés dans un but pronostique en ne prescrivant que ceux qui apportent une information pertinente et non redondante avec celle d'autres examens.

Tableau I – Différents intérêts de l'évaluation d'un pronostic.

1. La prédiction individuelle chez un malade et la connaissance de l'évolution d'une maladie.
2. L'adaptation d'un traitement à la gravité d'une maladie.
3. Dans l'élaboration du protocole d'un essai thérapeutique randomisé, la mise en œuvre éventuelle d'une stratification.
4. La limitation de la prescription d'examens complémentaires à visée pronostique à ceux qui sont nécessaires et suffisants.

L'évaluation d'un pronostic consiste à apprécier l'association entre, d'une part, un ou des facteurs susceptibles d'avoir un lien avec ce pronostic et appelés variables expliquantes ou covariables et, d'autre part, celui qui est appelé variable expliquée ou variable dépendante. Les variables expliquantes peuvent être des caractéristiques du sujet (âge, sexe, etc.) ou des facteurs de comorbidité (obésité, diabète, cardiopathie, etc.), mais surtout des facteurs présumés être liés à la gravité de la maladie. Les variables expliquées sont souvent des variables quantitatives (survie à un mois après le début de la maladie), mais plus souvent encore ce sont des données censurées (survie, apparition d'une récidive, etc.).
Dans toutes ces évaluations, la qualité d'un travail se juge d'abord sur la précision avec laquelle toutes ces variables ont été définies en privilégiant les critères objectifs d'appréciation sur les critères subjectifs.
Pour réaliser ces évaluations, les outils statistiques privilégiés sont les analyses unifactorielles, première étape des analyses multifactorielles, c'est-à-dire selon la nature des variables étudiées, la régression multiple, la régression logistique, le modèle de Cox et l'analyse discriminante (voir la deuxième partie de cet ouvrage).
Les analyses multifactorielles permettent de comprendre, au moins en partie, les liens qui existent entre des covariables et une variable expliquée. À partir de ces données, elles permettent de faire une prédiction. Pour ce faire, il convient d'abord d'affecter à chaque covariable un coefficient calculé à partir du logarithme de l'écart-type du logarithme

du risque relatif. Ensuite, le score de chacun des malades sur lesquels l'étude a porté est la somme des valeurs des covariables significatives affectées de leur coefficient. Des groupes de malades aussi identiques que possible sont ensuite constitués en fonction de leur score. Il convient enfin de valider les seuils ainsi proposés sur une ou des populations différentes de celle sur laquelle a porté l'étude initiale. À défaut, des validations internes sur la population initiale, comme celle dite de « Monte Carlo », permettent d'estimer la robustesse du score.

Exemple utilisant le modèle de Cox

1

Voici **un exemple** d'étude multifactorielle servant de point de départ à l'élaboration d'un score pronostique [1]. Cette étude concernait les malades qui avaient eu un cancer de l'œsophage réséqué chirurgicalement. La variable expliquée était la survie.
Dans un premier temps, 21 covariables ont été analysées. En analyse unifactorielle, neuf d'entre elles étaient liées à la survie (test du logrank).
Dans un deuxième temps, ces neuf covariables ont été incluses dans un modèle de Cox. Celui-ci a montré que seules quatre variables étaient indépendamment associées à un mauvais pronostic (tableau I).

Tableau I – Modèle de Cox. Résultats d'un travail sur les facteurs de pronostic de la survie de malades ayant eu un cancer de l'œsophage qui a été réséqué de façon apparemment curative [1].

Covariables	Coefficient de régression β	Écart-type	*p*	Risque relatif instantané
Âge > 65 ans	0,05	0,02	0,02	1,05
Classification ASA*	0,39	0,25	0,01	1,47
Infiltration pariétale	0,40	0,15	0,03	1,49
Envahissement ganglionnaire	0,38	0,19	0,01	1,46

* ASA : American Society of Anesthesiology. ASA est un score global de risque en quatre classes ordonnées qui tient compte des fonctions vitales d'un malade.

Ensuite, il convient d'affecter à chaque covariable indépendante un coefficient, calculé comme il a été indiqué, à partir du logarithme du risque relatif instantané (tableau II).

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Tableau II – Score pronostique de la survie de malades ayant eu un cancer de l'œsophage qui a été réséqué de façon apparemment curative.

Âge × 1,03 +	(0 si < 65 ans ; 1 si > 65 ans)
ASA* × 7,56 +	(*cette classification va de 0 à 4)
Infiltration pariétale × 7,85 +	1 si elle intéresse la sous-muqueuse
	2 si elle intéresse la musculeuse
	3 si elle dépasse la musculeuse
Envahissement ganglionnaire × 7,40	0 s'il n'y a pas de métastase ganglionnaire
	1 s'il y a métastase juxta-tumorale
	2 s'il y a métastase à distance de la tumeur

Les coefficients 0, 1, 2, 3 ne sont acceptables que si l'on admet l'hypothèse, qui est forte dans cet exemple, d'une évolution linéaire du risque relatif entre les classes de chaque variable.
Ces scores, comme celui-ci, sont en pratique difficiles à appliquer. Il est donc souhaitable de chercher à les simplifier. C'est ce qui a été fait dans notre exemple. L'âge qui avait une influence près de huit fois inférieure à celle des trois autres covariables a été supprimé. Les coefficients des trois autres covariables, étant assez proches les uns des autres ont été considérés comme égaux entre eux. On aboutit ainsi au score simplifié montré dans le tableau III.

Tableau III – Score pronostique simplifié de la survie de malades ayant eu un cancer de l'œsophage qui a été réséqué de façon apparemment curative.

ASA × 1 à 4
Infiltration pariétale × 1 à 3
Envahissement ganglionnaire × 0 à 2

La valeur de ce score simplifié pouvait aller de 2 à 9.
Les malades ont ensuite été regroupés en trois catégories d'effectifs peu différents les uns des autres :
– le groupe 1 constitué de malades qui avaient un score de 2 à 4 ;
– le groupe 2 constitué de malades qui avaient un score de 5 ou 6 ;
– le groupe 3 constitué de malades qui avaient un score de 6 à 9.
La figure 1 montre que les courbes de survie de ces trois groupes étaient bien séparées les unes des autres, ce qui suggère la qualité du modèle. Il aurait été souhaitable, ce qui n'a pas été fait, de valider ce score sur un autre ou sur d'autres échantillons de malades que celui à partir duquel il a été établi.

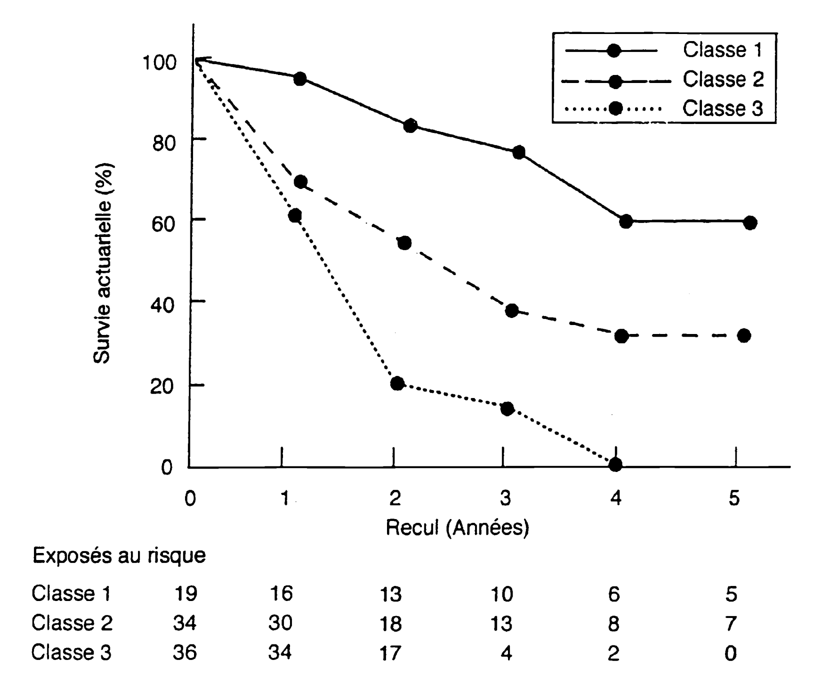


Fig. 1 – Courbes de survies de malades ayant eu un cancer de l'úsophage, réséqué, en fonction d'une classification en trois groupes à partir d'une étude multifactorielle utilisant le modèle de Cox.

Un autre exemple montre la combinaison de trois méthodes biostatistiques. Il concerne une classification pronostique des mélanomes malins primitifs. Dans une première étape, trois anatomopathologistes ont relu indépendamment les lames de 198 mélanomes malins de stade I [2]. Les critères de jugement ont été le type histologique, l'invasion dans le derme, l'épaisseur tumorale. Une étude de concordance a été faite pour chaque paramètre avec estimation des coefficients kappa. L'étude pronostique unifactorielle a été réalisée pour chaque pathologiste utilisant le test du logrank. Le critère de jugement du pronostic a été la durée de survie sans récidive. Les trois covariables ont ensuite été introduites dans un modèle de Cox dans un pas à pas ascendant. L'épaisseur de la tumeur a toujours été le premier paramètre sélectionné. Parmi les codages d'épaisseur utilisés, le codage optimal a été celui en deux classes, moins de 2 mm, et plus de 2 mm. En outre, cette classification avait abouti à des effectifs équilibrés pour cet échantillon et la concordance de jugement pour ces deux classes était très élevée (coefficient kappa > 0,70). L'étape suivante a été la validation du modèle sur un autre échantillon de 145 malades. Cette classification a permis de mettre en œuvre un essai randomisé. Ce travail est ainsi un modèle de démarche dans la rigueur scientifique d'un travail clinique.

1

Références

1. Pétrequin P, Huguier M, Lacaine F, Houry S (1997) Cancers de l'œsophage réséqués : modèle prédictif de survie. Gastroenterol Clin Biol 21: 12-6
2. Chastang C, Césarini JP, Beltzer-Garelli H, *et al.* (1984) Établissement d'une classification pronostique en 2 stades du mélanome malin primitif à partir d'une analyse multidimensionnelle et d'une étude de concordance. Rev Epidem et Santé Publ 32: 243-8

Exemple utilisant l'analyse discriminante

2

Cet exemple concerne des patients qui ont une cirrhose hépatique et qui ont fait une hémorragie digestive, le plus souvent due à une hypertension portale. La variable que l'on a cherché à expliquer est la mortalité à un mois [1].

Comme dans l'exemple précédent, il a d'abord été fait une analyse unifactorielle de chaque covariable que l'on pensait intéressant d'inclure dans le modèle. Le tableau I montre le coefficient de Mahalonobis et sa signification statistique.

Tableau I – Malades ayant une cirrhose hépatique et ayant eu une hémorragie digestive. Survie à un mois. Analyse unifactorielle.

Covariable	Coefficient de Mahalonobis	*p*	% de malades bien classés
Ascite	0,364	< 0,01	63
Bilirubinémie	0,303	< 0,01	70
Temps de Quick	0,286	< 0,01	67
Cause de l'hémorragie	0,158	< 0,05	61
Médicament gastro-agressif	0,154	< 0,05	58
Type de l'hémorragie	0,084	n.s.	48

Dans ce tableau, les covariables ont été classées par ordre décroissant du coefficient et de signification statistique. Les valeurs prédictives positives et négatives pour chaque covariable ont ensuite été calculées ainsi que le pourcentage de malades bien classés comme le montre la dernière colonne du tableau.

Selon le même principe, il a été fait ensuite une analyse multifactorielle en ajoutant à la covariable ascite les autres covariables une à une (tableau II).

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

Tableau II – Malades ayant une cirrhose hépatique et ayant eu une hémorragie digestive. Survie à un mois. Analyse multifactorielle.

Covariable	Coefficient de Mahalonobis	% de malades bien classés
Ascite	0,364	63
Ascite + bilirubinémie	0,587	72
Ascite + bilirubinémie + cause de l'hémorragie	0,764	72
Ascite + bilirubinémie + cause + Quick	0,864	75
17 covariables	1,710	79

Ce tableau montre qu'au-delà des quatre covariables les plus discriminantes, certes, plus on inclut de covariables, plus le pourcentage de malades bien classés augmente, mais ceci de plus en plus faiblement alors que le modèle devient de plus en plus complexe.
Notons encore que, dans cet exemple, les auteurs du travail ont utilisé l'analyse discriminante, mais ils auraient pu se servir de la régression logistique.

Grille de réalisation ou de lecture des analyses multifactorielles prédictives

1. L'objectif de l'analyse est pertinent sur le plan médical, c'est-à-dire que :

– le choix de la population étudiée ne doit pas aboutir à confirmer ce qui est déjà bien établi ;

– les covariables potentiellement importantes sont bien incluses dans l'étude.

2. Les définitions fondamentales sont clairement précisées :

– la population sur laquelle l'étude a porté ;

– les covariables expliquantes incluses dans le modèle sont bien définies ;

– il en est de même de la variable expliquée.

3. Le choix du modèle est correct

– si la variable expliquée est quantitative : utilisation de la régression multiple ;

– si la variable expliquée est qualitative à deux classes : utilisation de la régression logistique ;

– si la variable expliquée est censurée : utilisation du modèle de Cox ;

– l'analyse discriminante dans certains cas.

4. L'analyse a d'abord comporté une étude unifactorielle pour sélectionner, en général en retenant un $p < 0{,}25$, les covariales proposées à **l'analyse multifactorielle**.

Remarque : le modèle ne devait inclure au maximum qu'une covariable par dix événements, deux pour 20, trois pour 30, etc.

5. Dans les meilleurs travaux, **les résultats ont été validés** sur un ou plusieurs échantillons de malades différents de ceux de l'échantillon initial. Sinon, la discussion des auteurs du travail devrait, au moins, atténuer la portée des résultats observés.

Référence

1. Poynard T, Chaput JC, Mary JY, *et al.* (1980) Analyse critique des facteurs liés à la mortalité au trentième jour dans les hémorragies digestives hautes du cirrhotique. Gastroenterol Clin Biol 4: 655-65

partie

Épidémiologie

7

Introduction

Étymologiquement et historiquement, l'épidémiologie est l'étude des épidémies de maladies transmissibles. Mais, dans son acception moderne, l'épidémiologie est l'étude de tout événement de santé et de situations d'intérêt sanitaire, des associations entre ces événements ou ces situations et de l'impact sur la population des expositions aux facteurs de protection ou de risque [1]. Ces facteurs peuvent dépendre de caractéristiques de l'individu comme le sexe, l'âge, des facteurs génétiques, des comportements individuels comme l'alcoolisme, le tabagisme, la surcharge pondérale ou l'obésité, ou de l'environnement comme la pollution atmosphérique, les nuisances sonores ou bien de protection comme le contrôle de la tension artérielle.

Plus récemment, certains auteurs ont parlé d'épidémiologie clinique à propos de l'évaluation diagnostique, thérapeutique ou pronostique, c'est-à-dire pour tout ce qui concerne l'évaluation médicale.

Nous retiendrons ici uniquement ce qui concerne l'épidémiologie liée à la santé publique. C'est la science de l'analyse de la santé au niveau des populations, basée sur une approche descriptive et surtout comparative ou analytique. Elle repose sur l'utilisation de statistiques, donc au recours au calcul de probabilités et au concept de risque.

Il existe en effet, deux aspects de l'épidémiologie :

– **l'épidémiologie descriptive** qui étudie dans des populations la distribution, l'évolution au cours des années ou dans des zones géographiques différentes de phénomènes de santé bactériens ou viraux, mais aussi la couverture vaccinale, les suicides, les maladies chroniques, le risque thérapeutique, etc. Le moyen utilisé est une enquête transversale (à une date donnée (*cross-sectional* en anglais) ou longitudinale (un suivi sur une période). On peut suivre la totalité de la population ou un échantillon représentatif de celle-ci ;
– **l'épidémiologie analytique** qui étudie les facteurs susceptibles de favoriser la survenue de maladies. Elle utilise comme principales méthodes les enquêtes cas-témoins et les enquêtes de cohorte

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

(exposés-non exposés) ; les autres méthodes, cas-cohorte, cas-témoins emboîtés dans la cohorte, cas-croisés, étant des modèles dérivés ou apparentés aux deux précédents (tableau I).

Tableau I – Les différents types d'enquêtes épidémiologiques.

L'épidémiologie descriptive :
– les études d'observation transversales ou longitudinales.
L'épidémiologie analytique ou prédictive :
– les études cas-témoins ;
– les études de cohorte, encore appelées exposés-non exposés.

Référence

1. Flahault A, Spira A (2012) La situation de l'épidémiologie en France en 2011. Bull Acad Natle Med (sous presse)

L'épidémiologie descriptive : les enquêtes transversales

1

L'épidémiologie descriptive est synonyme de connaissances des indicateurs de santé. Elle a pour champ d'action essentiel l'étude de la mortalité et de la morbidité (reconnue ou ressentie) dans une population. C'est, par exemple, la « photographie épidémiologique » prise un jour donné dans une population pour recenser les surcharges pondérales et les obésités.

Mesure de fréquence (ou de risque absolu)

En épidémiologie descriptive, les études transversales consistent à recueillir des observations à une date donnée, et les enquêtes longitudinales au cours d'une période donnée. La mesure de fréquence d'un état de santé utilisé est la prévalence (tableau I).

Tableau I – Mesures de fréquence (ou de risque absolu).

Prévalence : nombre ou proportion de personnes concernées par un événement dans une population donnée, à un moment donné.
Incidence : nombre ou proportion de nouveaux cas au cours d'une période donnée, dans une population donnée.

La **prévalence** est la proportion de personnes concernées par l'événement de santé dans une population donnée à un moment donné. Cette prévalence ponctuelle peut s'exprimer par le nombre de cas dans la population, mais aussi en rapportant ce nombre de cas à une population, c'est-à-dire en pourcentage. Au sens démographique, c'est une statistique qui mesure l'état d'une population à un moment donné. La prévalence peut être analysée globalement ou selon le sexe, les tranches d'âge, la région, etc. Par exemple, en France, la prévalence du cancer du côlon-rectum était d'environ 109 000 personnes en 2002, ce qui correspond à 1,6 ‰ dans la population. Il est encore possible de mesurer

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

la prévalence sur une période donnée. On parle alors de prévalence de période.

En revanche, l'**incidence** est le nombre de nouveaux cas recensés pendant une période de temps donnée. Le taux d'incidence rapporte ce nombre à la population à risque. Au sens démographique, c'est une statistique qui mesure l'évolution de l'état d'une population dans un intervalle de temps. Par exemple, en France, l'incidence annuelle des cancers du côlon était de 40 500 cas, avec chez l'homme 40 nouveaux cas pour 100 000 personnes et chez la femme de 25 pour 100 000 [1]. Le taux d'incidence peut être calculé en comptant la population à risque sous forme de personnes années : on parle alors de densité d'incidence. Par exemple, si 1 000 personnes à risque sont suivies pendant deux ans et que 28 sont devenues séro-positives au VIH, pendant cette période, la densité d'incidence est de 1,4 pour 100 personnes années.

Prévalence et incidence sont deux notions complémentaires, comme le sont la position et la vitesse. Par exemple en 1998, le taux d'incidence de l'infection à VIH en France baissait, probablement en partie grâce aux mesures de prévention, tandis que la prévalence continuait à augmenter en raison de la longue durée d'incubation de la maladie et de l'effet des traitements qui prolongent la survie des malades. De même, en 1977, la loi qui renforçait le *numerus clausus* des étudiants en médecine en deuxième année allait réduire l'incidence du nombre de médecins, tandis que la prévalence devait continuer à augmenter jusqu'en 2003 environ.

Certaines études épidémiologiques descriptives reposent sur des données recueillies en permanence grâce, notamment, aux registres de morbidité (cancers, malformations congénitales, etc.) aux déclarations obligatoires de certaines maladies transmissibles ou d'effets indésirables de médicaments ou sur des données spécifiques d'études transversales. Ces dernières peuvent porter sur l'ensemble de la population ou sur un échantillon, au mieux représentatif de la population, c'est-à-dire tiré au sort.

Ce peut être encore les données recueillies chaque jour par un réseau de « médecins sentinelles » pour connaître en temps réel le niveau épidémique et établir la cartographie de la grippe, des gastro-entérites ou de tout autre maladie dans une population. En France, le réseau des « médecins sentinelles » de l'INSERM représente environ 1 % des médecins généralistes. Ils sont bénévoles et volontaires pour rapporter par voie téléinformatique à l'Institut national de la santé et de la recherche médicale, selon des protocoles standardisés, les informations qu'ils observent dans leur pratique quotidienne.

Répétition des mesures de fréquence

Les études transversales peuvent faire l'objet de **comparaisons dans le temps** : ce sont les études de séries chronologiques encore appelées « études avant-après » lorsqu'elles évaluent une intervention. Ce renouvellement dans le temps des enquêtes sur un même problème de santé d'une population, concerne, par exemple, le suivi d'une couverture vaccinale dans une classe d'âge ou le taux d'infections nosocomiales. Ce type d'étude permet d'estimer l'efficacité de certaines politiques de santé publique : campagnes pour la vaccination, mesures hospitalières d'hygiène.

Ces études transversales peuvent aussi faire l'objet de **comparaisons entre des populations différentes ou des régions différentes** ; ce sont les « études dites ici-ailleurs ». Ces comparaisons doivent parfois être ajustées pour ne pas fausser leur interprétation. Ainsi, lorsque l'on compare des taux de mortalité ou de morbidité entre des régions, il est nécessaire de procéder à une standardisation de ces taux sur l'âge. Par exemple, une enquête de la Caisse nationale d'assurance maladie a montré que l'on faisait deux fois plus d'appendicectomies par 100 000 habitants dans la région Nord-Pas-de-Calais que dans la région Provence-Alpes-Côte d'Azur [2]. Cependant, dans le Nord de la France, la population est plus jeune que dans le Sud : cette différence d'âge contribue à la différence car, dans le Sud, de nombreuses personnes auront déjà subi une appendicectomie dans l'enfance ou au début de l'âge adulte. Il est dans ce cas possible de comparer les taux par classe d'âge, par exemple chez les 10-20 ans dans les deux régions ; ou de standardiser les taux en se rapportant à une population de même âge. Dans notre exemple, l'ajustement des taux d'appendicectomies en fonction des classes d'âge a montré que la différence entre les deux régions persistait à classes d'âge similaires suggérant que des comportements médicaux différents expliquaient les différences observées.

Références

1. Launoy G, Grosclaude P, Pienkowski P *et al.* (1992) Cancers digestifs en France. Comparaison de l'incidence dans 7 départements et estimation de l'incidence en France. Gastroenterol Clin Biol 16 :633-8
2. Caisse nationale d'assurance maladie (1992) L'activité chirurgicale dans les établissements de santé. Résultats médicaux nationaux. Paris, Caisse nationale d'assurance, Tome 1 : 177-8

L'épidémiologie analytique 2

L'épidémiologie analytique, ou explicative, a pour but d'étudier des facteurs susceptibles de favoriser la survenue de maladies. Théoriquement, le meilleur niveau de preuve pour atteindre cet objectif serait de faire un essai randomisé. En réalité, il ne serait pas acceptable, sur le plan éthique après tirage au sort, d'exposer une partie des sujets inclus dans une telle étude à un facteur qui serait potentiellement pathogène. Il est donc nécessaire de procéder différemment.
Pour ce faire, on dispose de deux principaux types d'enquêtes explicatives. Leurs objectifs sont un peu différents. Les enquêtes cas-témoins partent d'un échantillon de sujets atteints d'une maladie que l'on compare à une série témoin de sujets non atteints par cette maladie. Les enquêtes de cohorte diffèrent dans la mesure où l'on compare deux groupes de sujets que l'on va suivre dans le temps, les uns exposés à un facteur de risque potentiel, les autres non exposés.

Les enquêtes cas-témoins

La caractéristique essentielle de cette méthode est que l'on constitue deux groupes de sujets, d'un côté des malades (M+), et de l'autre des personnes non malades, appelées témoins, contrôles, ou référents (M–). Les sujets du groupe témoin devront provenir de la même population que les cas. Autrement dit, il faut que les sujets témoins, s'ils avaient été des cas, aient pu être inclus dans le groupe des cas. Lorsque ceci n'est pas réalisé, des biais substantiels peuvent apparaître dans l'analyse. C'est donc bien la constitution du groupe témoin qui est le plus difficile à réaliser dans une telle étude. Le but des études cas-témoins est de comparer dans chacun des deux groupes la fréquence des expositions antérieures (E+) ou l'absence d'exposition (E–) à des facteurs de risque présumés comme le montre le tableau I. Dans ce type d'étude, par construction, le recueil de l'information sur les expositions et les facteurs de risque est rétrospectif.

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

2

Tableau I – Principes et objectifs des enquêtes cas-témoins.

On part de deux groupes, l'un de malades (M+) et l'autre de sujets non malades (M–).

On cherche à savoir si la fréquence des expositions à un facteur de risque (E+) ou non (E–) est différente entre ces groupes.

Cela permet de dresser le tableau suivant :

	M+	**M–**
E+	a	b
E–	c	d

À partir duquel, il est possible de calculer les odds et les odds ratio :

– *Odds* chez les malades : a/b

– *Odds* chez les non malades : c/d

– *Odds ratio* : $\frac{a \times d}{b \times c}$

Ces *odds ratio* sont indépendants de la fréquence de la maladie dans la population étudiée. Ils donnent une bonne approximation du risque relatif quand la maladie est rare et que l'enquête n'est pas biaisée.

Objectifs

Les enquêtes cas-témoins peuvent répondre à deux objectifs un peu différents.

Ou bien elles sont destinées à étayer ce qui n'était qu'une hypothèse d'une liaison entre, par exemple, une exposition à certains anorexigènes et la survenue d'une hypertension artérielle pulmonaire et à quantifier la force d'association. Pour ce faire, on observe un groupe de sujets ayant une hypertension artérielle pulmonaire et un groupe de sujets aussi similaires que possible n'ayant pas d'hypertension artérielle pulmonaire. On compare ensuite entre ces deux groupes, les niveaux d'exposition antérieure aux anorexigènes[5].

Ou bien ces enquêtes sont *exploratoires* afin de tester un certain nombre d'hypothèses. C'est ainsi qu'un grand nombre de facteurs de risques potentiels du carcinome vaginal chez la jeune fille avaient été recherchés et ont permis de suspecter la prise de Distilbène® par la mère. L'utilité décisionnelle de ces enquêtes cas-témoins peut être illustrée

5 Cet exemple avait été donné dans notre ouvrage « Biostatistiques au quotidien » publié par Elsevier, Paris en 2000.

par un autre exemple [1]. Des médecins généralistes ont interrogé 500 de leurs malades ayant consulté au lendemain des fêtes de fin d'année 1995 pour une gastro-entérite. Trente pour cent de ces malades avaient consommé des huîtres dans les dix jours qui avaient précédé les symptômes. Il était donc tentant de penser que les huîtres étaient responsables d'une grande partie de ces gastro-entérites. En réalité, il était nécessaire, avant de tirer une telle conclusion avec les conséquences que cela aurait pu avoir en matière économique, notamment pour les ostréiculteurs, d'observer un groupe de sujets similaires, mais qui n'avaient pas eu de gastro-entérite. Cette étude a été réalisée par les médecins du « réseau sentinelle ». Il s'est avéré que la consommation d'huîtres chez les sujets témoins était la même que celle des malades qui avaient eu une gastro-entérite. On ne pouvait donc probablement pas incriminer de façon déterminante les huîtres, dans cette épidémie hivernale de gastro-entérite. Le fait que 30 % des malades atteints de gastro-entérite avaient consommé des huîtres s'expliquait par le fait, qu'à cette période de l'année, beaucoup de Français mangent des huîtres. Les résultats de cette enquête suggéraient alors de chercher d'autres causes à l'origine des gastro-entérites qui avaient été observées. De fait, en 1999, ce même réseau de médecins a permis d'identifier la cause d'au moins la moitié des cas de gastro-entérites survenant lors de ces épidémies hivernales. Il s'agissait de familles de virus entériques, principalement de *Calicivirus* et de *Rotavirus* [2].
La mesure de la force d'association dans une étude cas-témoins est l'*odd ratio* dont le calcul est indépendant de la fréquence de la maladie (tableau I). En revanche, il n'est pas possible de calculer directement le risque relatif qui, lui, dépend de la fréquence de la maladie.

Principaux avantages et inconvénients

Ces études cas-témoins ont l'avantage d'être relativement rapides à réaliser, car elles ne nécessitent pas le suivi des personnes. Leur coût est beaucoup moins élevé que celui des enquêtes prospectives (cohorte) ou des essais thérapeutiques. Elles permettent d'explorer plusieurs hypothèses simultanément, notamment par des analyses multifactorielles utilisant la régression logistique.
Elles ont cependant des inconvénients. Tout d'abord elles ne permettent pas de connaître l'incidence d'une maladie puisque le nombre de cas et de témoins est fixé de manière arbitraire par l'investigateur. De plus, s'il est assez aisé de constituer un groupe de malades, il est habituellement beaucoup plus difficile de trouver et de choisir des sujets témoins appropriés. Enfin, de nombreux biais peuvent être

présents dans ces analyses : biais de mémorisation : il est probable que les sujets malades se remémorent plus d'expositions suspectes qu'un sujet témoin indemne de la maladie ; biais de participation : imaginons que dans une étude sur les tumeurs cérébrales et le téléphone portable, les cas soient plus susceptibles de participer que les témoins s'ils ont utilisé un téléphone portable ; biais de sélection : lorsque la population de témoins représente imparfaitement la population à risque, par exemple si l'on recrute des témoins à l'hôpital plutôt qu'en population générale ; biais de classement lorsque par exemple la mesure de l'exposition sera réalisée avec plus d'erreur chez les témoins, biais de confusion lorsqu'une association retrouvée (par exemple consommation d'alcool et cancer du poumon) n'est pas causale mais s'explique par une tierce variable (ici le tabagisme), etc.
Une partie importante de l'interprétation de ces études est donc de vérifier que l'impact des biais et facteurs de confusion n'infirme pas la conclusion. Pour ces différentes raisons, les enquêtes cas-témoins sont surtout utilisées lorsqu'on ne peut pas réaliser de cohorte, c'est-à-dire pour des maladies rares ou dont le délai de survenue après une exposition présumée à un facteur de risque responsable est très long. Elles sont encore volontiers utilisées lorsque l'on a besoin d'une réponse rapide pour faire face à un risque sanitaire, par exemple lors d'une épidémie de listériose.

Les enquêtes de cohortes exposés non-exposés

La cohorte romaine était constituée de 600 soldats. Le mot cohorte a ensuite été repris pour évoquer d'importants effectifs de populations en déplacement. En épidémiologie, une cohorte est l'étude dite « longitudinale » d'un échantillon de sujets initialement non malades, mais les uns exposés à un (ou des) risque(s), les autres non exposés. L'objectif est de mesurer la survenue d'événements de santé (maladie, décès) au sein de cet ensemble de sujets, puis de comparer l'évolution du nombre de nouveaux cas entre sujets exposés et non exposés. Ces études se différencient bien des enquêtes cas témoins qui cherchent, comme on vient de le voir, rétrospectivement une exposition à des facteurs de risque hypothétiques auprès de groupes de malades et de non malades.

Moyens d'estimation

Les enquêtes de cohortes comparatives permettent d'estimer directement le risque relatif associé à l'exposition. Lorsque plusieurs facteurs de risque potentiel sont étudiés, il est possible et souhaitable de faire des analyses multifactorielles.

Objectifs

Il existe deux types d'enquêtes exposés-non exposés.
Certaines **cohortes** sont dites « **historiques** » ou « **rétrospectives** ». Elles reposent sur le recueil rétrospectif des données en se basant sur des fichiers déjà constitués pour une autre raison. Par exemple, des cohortes historiques ont été utilisées pour étudier l'association entre l'exposition professionnelle à l'amiante et la survenue d'un mésothéliome. Il a suffi de reprendre, en milieu professionnel, les comptes rendus des visites en médecine du travail et les radiographies thoraciques d'un groupe de travailleurs ayant eu une profession les exposant à l'amiante et d'un groupe de travailleurs de même âge, de même sexe et ayant la même consommation de tabac et d'alcool, mais non exposés à l'amiante. Cette comparaison a montré que l'exposition professionnelle à l'amiante constituait bien un facteur de risque de mésothéliome. Dans les pays où la réglementation l'autorise, le croisement de fichiers nominatifs permet de réaliser des cohortes historiques « virtuelles » ou « électroniques » qui ouvrent de nombreuses et utiles possibilités. Ainsi, au Canada, le croisement de fichiers de patients qui ont reçu une prescription de médicaments et de fichiers de patients hospitalisés pour des réactions indésirables suspectes d'être dues au même médicament, a montré le rôle des β2 mimétiques utilisés seuls, sans corticoïdes, comme facteur de risque d'état de mal asthmatique [3].
D'autres études de cohortes comparatives reposent sur le **recueil prospectif** de deux cohortes parallèles, l'une exposée à un facteur de risque (par exemple le tabagisme), et l'autre non exposée. Les sujets sont alors suivis plusieurs mois, voire plusieurs années, pendant lesquelles on collige la survenue d'événements présumés liés à l'exposition (par exemple les broncho-pneumopathies et le cancer du poumon). Il convient de remarquer que l'essai randomisé est une forme particulière de cohorte comparative avec la particularité que les deux cohortes sont déterminées par le tirage au sort. Les cohortes comparatives sont, après les essais randomisés, les méthodes de comparaison les moins sujettes à des biais.

Les études avec recueil prospectif des données sont surtout intéressantes et utiles lorsque le délai pressenti entre l'exposition aux facteurs de risque et l'apparition d'une maladie est relativement bref et que l'incidence de la maladie est élevée.

En revanche, les délais d'observation peuvent être trop longs. Par exemple, pour l'exposition à l'amiante, il aurait fallu attendre 30 ans de recueil prospectif des données avant que les mésothéliomes apparaissent pour se rendre compte que l'amiante augmentait leur risque de survenue.

Au pire, lorsque l'exposition est assez fréquente et distribuée de manière hétérogène dans la population et que les événements (maladie, décès) sont rares ou surviennent tardivement après l'exposition au facteur de risque potentiel, l'étude de cohortes est pratiquement impossible à réaliser. Par exemple, pour avoir 95 % de chances de détecter un cas d'une maladie dont la fréquence serait de 1/10 000, il faudrait suivre 30 000 personnes. La constitution d'une telle cohorte présenterait des difficultés logistiques difficilement surmontables et un coût considérable alors que quelques cas et leurs témoins permettent parfois de reconnaître d'éventuelles relations entre l'exposition à un facteur de risque et l'apparition d'une maladie.

Le tableau II montre les données qui peuvent guider le choix d'une méthode épidémiologique.

Tableau II – Choix d'un type d'étude épidémiologique en fonction des possibilités.

1. Maladies rares ou dont le délai de survenue après l'exposition au facteur de risque est très long ou bien que l'on a besoin d'une réponse rapide, par exemple en cas d'alerte sanitaire :

→ Étude cas-témoins.

2. Maladies assez fréquentes, mais dont le délai de survenue après l'exposition au facteur de risque est long :

→ Cohortes « historiques ».

3. Maladies dont l'incidence est élevée et le délai de survenue après l'exposition au facteur de risque est relativement bref :

→ Cohortes avec recueil prospectif des données.

Cette dernière méthode est celle qui est le moins sujette à des biais. Au mieux, l'allocation des sujets dans le groupe exposé ou non exposé est réalisée par tirage au sort : c'est l'essai randomisé, mais qui, en dehors d'une perspective de prévention, est peu praticable en épidémiologie pour des raisons éthiques.

Les biais

Le groupe témoin

Dans toutes les enquêtes épidémiologiques, le choix pertinent du groupe témoin, c'est-à-dire dans les enquêtes cas-témoins les non-malades et dans les enquêtes exposés-non exposés, les non-exposés, est fondamental. En effet, lorsqu'un tirage au sort n'est pas possible comme dans un essai thérapeutique par exemple, il est indispensable que les échantillons de groupes témoins soient le plus représentatifs possible de la population afin de limiter le risque de biais lié à une sélection des sujets au moment de leurs recrutements.
D'autres biais, lors de l'analyse des résultats, sont le rejet des cas parce que des données sont partiellement manquantes ou encore, dans le suivi d'une cohorte, le fait que des sujets ont été perdus de vue. Des techniques permettent de limiter les conséquences de ces biais comme l'analyse actuarielle des résultats (cf. les variables censurées p. 37).

Des biais peuvent être liés aux erreurs de mesures

Des erreurs peuvent être liées à un questionnaire mal conçu au départ, à des appareils de mesure insuffisamment précis, aux observateurs eux-mêmes, etc. La validation des dossiers d'inclusion dans l'étude, le contrôle du suivi et des résultats observés, permettent de quantifier l'impact de ces erreurs sur les résultats, voire de les corriger.

Les facteurs de confusion

Ces facteurs ont été évoqués à propos des études mutifactorielles. Il y a facteur de confusion lorsque, de façon à la fois simultanée et indépendante, l'exposition et l'événement de santé sont influencés par un facteur extérieur qui n'a pas été pris en compte. Un facteur de confusion est le fait d'une association réelle, mais qui n'est pas causale pour autant. On pourrait donc dire qu'il ne s'agit pas d'un biais *stricto sensu.*

Remarques

Effectifs

Dans les enquêtes cas-témoin, comme dans les enquêtes de cohorte, le calcul du nombre de sujets nécessaire se fait sur la base de la comparaison de deux pourcentages : exposés chez les cas contre exposés chez les témoins dans l'étude cas-témoin ; incidence chez les exposés contre incidence chez les non-exposés dans la cohorte.

Dans les enquêtes cas-témoins, le facteur limitant est généralement le nombre de cas. Sur le plan de la puissance statistique, on peut augmenter la puissance en incluant plusieurs sujets témoins pour chaque cas. Néanmoins, au-delà de quatre à cinq sujets témoins par cas, ce gain de puissance devient négligeable. Lorsque les différences d'exposition sont majeures, ce type d'étude peut être très efficace. Il avait suffi de huit cas et de 32 témoins pour montrer que la prise de stilbestrol chez les femmes enceintes favorisait la survenue d'un cancer du vagin chez leurs filles [4]. Ce cancer apparaît dans l'adolescence et cette étude a fait proscrire formellement le stilbestrol pendant la grossesse.

Dans l'enquête de cohorte, on pourra choisir d'inclure un échantillon représentatif de la population pour l'exposition. Cependant, pour maximiser la puissance à nombre de participants fixé, il sera mieux d'inclure autant d'exposés que de non-exposés. Lorsque cela n'est pas possible, on peut déséquilibrer les groupes, mais il faudra alors augmenter les effectifs pour garder la même puissance.

Causalité

Une fois les données recueillies, la comparaison des résultats entre les deux groupes cas-témoins, exposés-non exposés doit faire intervenir des tests statistiques comme dans toute comparaison. Mais, en dehors des essais randomisés, des différences statistiquement significatives entre un facteur d'exposition et une maladie ne signifient pas nécessairement qu'il y ait un lien de causalité entre eux.

Rappelons que pour suspecter une lésion causale, il ne suffit pas qu'il y ait des différences statistiquement significatives entre une exposition ou non à un risque et une maladie, mais qu'il faut que des conditions supplémentaires soient réunies. On donne la liste ci-dessous des critères dus à Bradford-Hill, qui sont couramment utilisés, sans pour autant apporter une garantie de causalité.

1. Le risque de maladie doit être statistiquement plus élevé lorsque l'on est exposé au facteur de risque considéré que si l'on ne l'est pas.
2. Les études épidémiologiques doivent ensuite établir des associations entre les facteurs d'exposition et le risque de survenue de maladie, dont la force est estimée par le risque relatif.
3. Il convient encore que l'association soit confirmée dans plusieurs études utilisant, de ce fait, des méthodes parfois un peu différentes les unes des autres, portant sur des populations différentes.
4. Les relations entre l'intensité du facteur de risque, c'est-à-dire la durée et la dose d'exposition (par exemple pour le tabagisme, le nombre de paquets années), sont des arguments supplémentaires [5].
5. L'exposition au facteur de risque doit précéder la survenue de la maladie.
6. Ajoutons enfin les arguments expérimentaux et la plausibilité biologique.

Aucune de ces données ne peut, à elle seule, apporter une preuve indiscutable de causalité, mais aussi, aucune ne doit être considérée comme un critère indispensable pour affirmer la causalité, sauf bien entendu la temporalité qui est la séquence dans le temps : exposition – survenue de la maladie.

Risques relatifs et risques attribuables

Une difficulté supplémentaire en épidémiologie est que, même si une relation causale existe entre un facteur de risque et la survenue d'une maladie, toute survenue de la maladie ne sera pas due à ce facteur, et toute exposition ne déclenchera pas la maladie. Les relations épidémiologiques sont avant tout de nature probabiliste. Chacun sait qu'un cancer du poumon peut se voir chez un malade qui n'a jamais fumé, mais que le tabagisme augmente ce risque. Une étude avait montré qu'entre 50 ans et 69 ans, le fumeur multipliait par 78 le risque qu'il avait de mourir d'un cancer du poumon dans les quatre années suivantes [6]. En termes de santé publique, il est important de connaître l'impact du facteur de risque à l'échelle d'une population, par exemple sous forme du nombre de cas attribuables à ce facteur.

La fraction du risque attribuable est fonction du risque relatif, mais aussi de la proportion de sujets exposés au facteur de risque dans la population. C'est ainsi qu'un facteur de risque peut entraîner un risque relatif très élevé d'une maladie, mais seulement un très petit nombre de cas si peu de personnes sont exposées à ce facteur de risque. Inversement, si un facteur de risque n'a qu'un faible pouvoir pathogène, mais que ce facteur est très répandu, il pourra générer un grand nombre

d'affections dans la population. Par exemple, le formaldéhyde augmente le risque relatif de cancer du nasopharynx de 1,5 à 2. Ce produit est utilisé dans de nombreux objets courants, mais il n'augmente le risque de ce cancer qui ne s'observe qu'entre 2 et 5 cas/million d'habitants qu'à des niveaux importants d'exposition. La relation de causalité qui a été montrée entre une exposition importante au formaldéhyde et le cancer du nasopharynx ne justifie donc pas, en termes de fraction du risque attribuable, des mesures de santé publique majeures dans la population générale. En revanche, il est très important de réduire le niveau d'exposition des travailleurs exposés à des niveaux élevés [7].

Imputabilité

Certaines maladies sont facilement imputables à une cause parce qu'elles sont particulièrement graves, voire mortelles et qu'elles succèdent à brève échéance au risque : affection virale fulgurante, exposition à des polluants à très forte dose, etc. Mais la grande majorité des maladies ont des causes multiples. Il devient alors très difficile, voire impossible, de montrer qu'une maladie est imputable à un facteur déterminé. Cette impossibilité pose des questions sociétales, notamment en matière d'indemnisation des personnes atteintes par une maladie susceptible d'être favorisée par un facteur de risque environnemental ou médicamenteux par exemple.

Références

1. Letrilliart L, Desenclos JC, Flahault A (1997) Risk factors for outbreak of acute diarrhea in France: case-control study. Br Med J 315: 1645-9
2. Brachet R, Etienney I, Flahault A, *et al.* (1999) Gastro-entérites hivernales : *Calicivirus* et *Rotavirus* ont été les deux familles de virus les plus fréquemment identifiées. Le Quotidien du médecin
3. Spitzer WO, Suisssa S, Ernst P, *et al.* (1992) The use of β-antagonist and the risk of death and near death from asthma. New Engl J Med 326: 501-6
4. Herbst AL, Ulfelder H, Poskanzer DC (1971) Adenocarcinoma of the vagin. Associaition of maternal stilbestrol therapy with tumor appearance in young women. N Engl J Med 284: 878-81
5. Huguier M (1976) Le tabac : risques calculables. Le Concours Médical 98: 7291-3
6. Hammond EL, Horn D (1958) Smoking and death rates. Report on forty-four months of follow-up of 187 783 men. JAMA I. Total mortality 166: 1159-72
7. Flahault A, Spira A (2011) La situation de l'épidémiologie en France. Bull Acad Natle Med (sous presse)

Prévention et dépistage 3

Prévention

Les enquêtes épidémiologiques trouvent un maximum d'intérêt lorsqu'elles ouvrent la possibilité de la mise en œuvre de mesures de prévention, de dépistage ou de maîtrise des risques. Encore faut-il prouver qu'il y a non seulement une association statistique, mais encore un lien de causalité entre un facteur de risque et une maladie.
Une fois un facteur de risque reconnu, toute politique de prévention devrait être assortie d'indicateurs qui permettent d'évaluer son efficacité, par exemple, le suivi de mesures d'impact, comme la mesure du risque attribuable avant et après l'intervention à visée préventive. Ainsi, en France, les campagnes menées entre 1985 et 2003 pour diminuer le tabagisme ont été très efficaces parmi les cadres chez lesquels la proportion de ceux qui consommaient régulièrement des cigarettes a diminué de 45 % à 2 %, alors que chez les ouvriers, cette diminution a été moindre, passant de 56 % à 49 % chez les hommes et qu'elle a même augmenté dans cette catégorie socioprofessionnelle chez les femmes de 19 % à 31 %.

Dépistage

Le dépistage, au sens strict, est le diagnostic d'une maladie avant l'apparition de symptômes ou de signes cliniques. L'hypothèse qui justifie le dépistage est que le traitement plus précoce de la maladie qu'il permet, améliore le pronostic ou réduit le risque que fait courir la maladie.
L'évaluation des moyens de dépistage dont on dispose se fait, comme pour le diagnostic en termes de sensibilité, de spécificité et de valeurs prédictives. De façon générale, le dépistage de masse utilise de première intention un examen sensible, sans contreparties médicales, peu

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

onéreux, mais souvent peu spécifique comme l'Hémocult® ou mieux l'immunologie pour dépister la présence de sang dans les selles. Dans un deuxième temps, dans le sous-groupe ainsi sélectionné, on réalise des examens plus spécifiques, par exemple une coloscopie pour reconnaître la présence éventuelle d'une tumeur colique.

En divisant le coût total du dépistage par le nombre de cas dépistés, il est possible de calculer le coût unitaire du dépistage.

Il convient encore de s'assurer que le dépistage a un intérêt décisionnel. Par exemple, l'utilité d'un dépistage du cancer de la prostate audelà de 70 ans peut être discutée, si l'histoire naturelle de ce cancer, à cet âge, montre qu'il évolue lentement, reste longtemps pas plus symptomatique qu'un adénome et que la probabilité de mourir d'une autre cause est plus élevée que celle de mourir des conséquences directes de ce cancer de la prostate.

En fait, si le bénéfice d'un dépistage n'est pas évident ou contesté, un essai randomisé est indiqué comparant la survie d'un sous-groupe de sujets qui ont eu un dépistage et le traitement éventuel qui en découle s'il est positif et un groupe témoin non dépisté. Ce dépistage peut s'étendre à la surveillance d'un malade guéri, mais susceptible de faire une récidive de sa maladie. Dans ces cas, la surveillance a pour objectif de dépister une récidive qui, reconnue plus tôt, avant d'être devenue symptomatique, serait plus facile à traiter. Malgré la logique apparente de ce type de raisonnement, des essais randomisés sont justifiés pour s'en assurer. Or plusieurs de ces essais, contrairement à ce que l'on pouvait logiquement espérer, n'ont pas montré d'amélioration de la survie grâce au dépistage. Ainsi, un essai randomisé a été réalisé chez 325 malades qui avaient eu une résection apparemment complète d'un cancer colorectal [1]. Deux protocoles de surveillance ont été comparés, l'un par de simples visites médicales, l'autre intensif avec plusieurs examens complémentaires susceptibles de dépister une récidive. Le critère de jugement principal a été la survie. Il n'a pas été observé de différence entre les deux groupes, ce qui a remis en cause l'utilité de programmes de surveillance intensive et onéreuse dans ces cas.

Référence

1. Shoemaker D, Black R, Gilles L, Toouli J(1998) Yearly colonoscopy, liver CT and chest radiography do not influence 5-year survival of colorectal cancer patients. Gastroenterology 114: 7-14

Épidémiologie théorique

4

L'épidémiologie théorique [1] repose sur la modélisation des événements de santé dans des populations. La théorie mathématique des épidémies repose sur le paradigme de la contagion interhumaine. Celle-ci peut être directe en cas de grippe par exemple. Il est alors possible de prévoir le nombre de cas dans une population en utilisant le taux de reproduction de base, c'est-à-dire le nombre de cas secondaires infectés directement par un cas index. Si la contagion se fait par l'intermédiaire d'un hôte vecteur, moustique par exemple en cas de paludisme, le même type d'approche est possible en prenant en compte le trajet particulier du pathogène.

Il est ensuite possible de reconstituer, à l'aide d'équations mathématiques ou de simulations sur ordinateur, des dynamiques de transmission d'agents infectieux, bactériens ou viraux. La théorie mathématique fournit des trajectoires épidémiques dans le temps et dans l'espace. Lorsqu'elle repose sur des observations initiales, elle autorise des simulations de scénarios qu'elle permet d'évaluer. Il est encore possible de proposer des prédictions qui aident des choix politiques de santé publique, par exemple, la couverture vaccinale minimale à atteindre pour obtenir un niveau de protection suffisant dans la population, l'âge optimal de la vaccination, l'impact de fermeture des locaux scolaires, de certains lieux publics ou des aéroports. Ce sont ainsi des modèles mathématiques qui ont montré l'inutilité de la fermeture des aéroports dans le cas de la grippe H1N1, ce qui aurait eu de très lourdes conséquences économiques et sociales pour un gain très faible, voire nul, sur le plan de la santé des populations.

L'exemple de l'épidémie de grippe H1N1 montre cependant la difficulté de ces prévisions initiales. Les premiers modèles ont reposé sur les observations de cas survenus au Mexique en avril 2009. Ceux publiés en France par l'Institut de veille sanitaire, en octobre la même année, estimaient que le nombre de décès pourrait se situer entre 3 000 et 90 000, ce qui était une fourchette très large. En définitive,

Michel Huguier et Pierre-Yves Boëlle, *Biostatistiques pour le clinicien*
ISBN : 978-2-8178-0463-7, © Springer-Verlag Paris 2013

le décompte fin 2010 a montré 312 décès. Tous les autres organismes de veille sanitaire dans les pays développés ont également surestimé le risque, en basant leurs projections notamment sur la plus sévère des pandémies passées (celle de 1918), alors que l'on a maintenant observé que sur les 5 dernières pandémies grippales, 4 avaient eu un impact modéré en nombre, quoiqu'important car touchant plutôt des individus jeunes. Bien entendu, les prévisions peuvent s'affiner avec le temps et se rapprocher alors de la réalité.
En conclusion, autant l'estimation précoce de l'importance d'une épidémie est difficile, autant l'apport de la modélisation mathématique comme aide à la décision d'une politique de contrôle et de prévention est souvent précieux.

Référence

1. Flahault A, Spira A (2011) La situation de l'épidémiologie en France. Bull Acad Natle Med (sous presse)

Les logiciels de biostatistiques

Nous donnons, à titre indicatif, une liste de logiciels statistiques couramment utilisés. Dans tous ces logiciels, les avantages et les inconvénients sont très subjectifs. Dans un traitement statistique, la phase qui prend généralement le plus de temps est celle de la saisie et du nettoyage des données, bien plus que celle de l'analyse proprement dite. Nous attirons donc l'attention sur la nécessité d'une conception ergonomique du cahier d'observation, afin de faciliter ces étapes.
Selon le volume d'information à saisir, l'organisation de la saisie elle-même, un tableur (par exemple Open Office Calc, Microsoft EXCEL™) pourra être suffisant ; l'utilisation de bases de données s'avérera nécessaire en cas de gros volumes et de questionnaires dépendants (Open Office BASE, Microsoft ACCESS™ ou SQL Server™, MySQL™Ö). De plus en plus, les logiciels de statistiques permettent également de créer des outils de saisie. On attirera également l'attention sur la disponibilité de solution de saisie à distance par exemple avec LIMESurvey qui permet la saisie par Internet. Finalement, toute base de données doit être déclarée à la CNIL et recevoir une autorisation.

Ci-après, une liste non exhaustive des logiciels utilisables est donnée par ordre alphabétique. À noter que d'autres logiciels, non spécifiques, offrent des capacités statistiques : certains tableurs ainsi que certains logiciels tournés vers la réalisation de graphiques. Pour des analyses simples, le choix du logiciel n'est pas essentiel, tous les logiciels proposant au minimum les tests usuels. Le choix devra être fait avant tout en fonction des ressources disponibles, de l'appétence envers l'univers informatique, et des possibilités locales de soutien !
Biostatgv : Le site internet Biostatgv (http://www.u707.jussieu.fr/biostatgv) permet de réaliser la plupart des tests classiques. Les calculs sont faits dans le logiciel R (voir ci-après). Il s'agit donc d'une solution simple pour effectuer des analyses descriptives, ainsi que des tests de différence, d'association. Il est également possible de calculer la taille des essais ou nombre de sujets.
R : Le logiciel R est un logiciel libre, gratuit, disponible sur le site CRAN (http://cran.r-project.org). Ce logiciel est très utilisé dans le milieu académique. Il peut être enrichi facilement par un système de

bibliothèques qui apporte des fonctionnalités supplémentaires. L'interaction avec le logiciel prend la forme de « scripts » ou programmes qui vont indiquer les traitements ou transformations que l'on souhaite appliquer aux données. Il est gratuit, extensible et les méthodes modernes sont rapidement disponibles. Il existe des extensions (RCommander) dont le but est de rendre l'utilisation plus ergonomique, la plupart des commandes étant réalisées au clavier.

S-plus™ : Il s'agit d'un logiciel commercial. S-plus est très semblable, dans ses capacités, au logiciel R décrit plus haut. Il bénéficie d'une interface plus conviviale, permettant de réaliser un bon nombre d'opérations en utilisant la souris (lecture des données, analyses standards), cependant le mode d'interaction privilégié reste le clavier. Il bénéficie d'une compatibilité très bonne avec le logiciel R qui permet de bénéficier des librairies développées pour ce dernier. L'accès rapide aux innovations statistiques est donc possible par le biais de librairies additionnelles.

SAS™ : le logiciel SAS est un logiciel commercial, édité par la compagnie SAS. SAS implémente une très grande variété de méthodes statistiques. Il est particulièrement performant dans le traitement de gros volumes de données, et très utilisé dans le milieu industriel. L'interaction avec le logiciel a lieu principalement sous la forme de scripts ou de programmes qui décrivent les traitements ou les transformations que l'on souhaite appliquer aux données.

SPSS™ : Il s'agit d'un logiciel commercial, édité par IBM™. L'abord de SPSS ressemble à un tableur, ce qui rendra le logiciel familier et diminuera la courbe d'apprentissage initiale. SPSS permet de réaliser les tests classiques rencontrés en recherche clinique et en épidémiologie. L'interaction peut aussi avoir lieu par le biais de « scripts » ou petits programmes entrés au clavier.

STATA™ : Il s'agit d'un logiciel commercial, édité par StataCorp™. Ce logiciel propose un choix important de méthodes classiques et modernes. Il met également en avant la possibilité de créer de nouvelles analyses par la programmation. L'interaction a lieu principalement par le clavier, plus que par la souris.

Quelques notations en biostatistiques

On utilise généralement les lettres majuscules (*X*, *Y*, *Z*, *P*, etc.) pour désigner des variables aléatoires. Par exemple, la proportion *P* de métastases hépatiques chez les malades atteints d'un cancer est une variable aléatoire (quantitative) qui prend des valeurs *p*0 différentes dans chaque échantillon de malades observés. De même, le taux de cholestérol (T*c*) dans une population française, etc. Si l'on mesure ces variables, par exemple sur un groupe de sujets ou de malades ou sur une série d'expériences (c'est-à-dire sur un échantillon), la valeur que prend la variable s'écrit en minuscules : *s* est une mesure (ou réalisation) de *X*, *y* est une mesure ou réalisation de *Y*, etc.
La notation est différente selon que l'on indique par une lettre grecque ce que serait la « vraie valeur » qui est rarement connue, par exemple le pourcentage π de nouveau-nés de sexe masculin dans la population française qui est de 51,5 % ou bien par une lettre latine minuscule, la valeur mesurée, dite estimée dans un groupe de personnes que l'on peut considérer être des échantillons de cette population (tableau I).

Tableau I – Notation des valeurs selon qu'elles sont réelles ou une estimation sur un échantillon.

	« Vraie valeur »	**Valeur estimée sur un échantillon**
Probabilité	π	p
Complément de la probabilité inverse	$1 - \pi$	$q = 1 - p$
Moyenne	μ	m
Variance	σ^2	s^2
Écart-type	σ	s
Coefficient de corrélation	ρ	r
Coefficient de concordance	κ	k

Les grandes lois de probabilité se notent avec des lettres majuscules cursives avec, entre parenthèses, les paramètres de la loi correspondante :
– loi normale $N(\mu, \sigma^2)$;

– loi binomiale $B\ (n, \pi)$;
– loi de Poisson $P\ (\lambda)$.

D'autres notations ont été utilisées :
– $\Sigma\ x$ représente la somme des valeurs x de l'échantillon ;
– i représente un individu, i^1, le premier de l'échantillon, i^2 le second, etc. ;
– x_i est la variable mesurée chez l'individu i correspondant ;
– N est la taille de l'effectif de l'échantillon étudié ;
– C_4^2 représente une combinatoire, c'est-à-dire dans cet exemple le nombre de façons de classer ou de ranger deux sujets parmi une liste de quatre ;
– s_A^2 signifie la variance (s^2) de l'échantillon A ;
– s_A signifie l'écart-type (s) de l'échantillon A ;
– $|X|$ veut dire la valeur absolue de X (c'est-à-dire que cette valeur soit $+\ X$ ou $-\ X$) ;
– ! est une factorielle, c'est-à-dire le produit dont les facteurs sont tous les entiers successifs égaux ou inférieurs à un nombre donné. Par exemple : $!\ 4 = 4 \times 3 \times 2 \times 1 = 24$.

Lexique

Les * renvoient à un autre mot.
Les mots entre [ ] sont les termes anglais correspondants.

A

Actuarielle (méthode) [*actuarial method*]
Méthode d'estimation adaptée aux variables censurées* (survie, récidive, etc.). Elle repose sur le principe des probabilités conditionnelles*. Les taux de survie sont évalués à intervalles réguliers, par exemple tous les 6 mois, tous les ans.

Ajustement [*adjustment*]
Ceci consiste à prendre en compte l'influence d'une tierce variable dans la mesure de la corrélation entre deux variables d'intérêt. Le but est de déterminer si la corrélation persiste lors de cet ajustement. Il est la base des études multifactorielles*.

Aléatoire (variable) [*random*]
Des variables sont dites aléatoires lorsque leur valeur dépend de l'individu sur lequel elles sont mesurées. Elles se différencient ainsi des constantes, plus souvent présentes en physique ou en mathématiques.

Alpha, α (risque)
Désigne le risque de première espèce*.

Ambivalence (clause d') [*ambivalence clause*]
Dans un essai randomisé* les sujets inclus doivent pouvoir recevoir l'une ou l'autre des interventions que l'on cherche à comparer.

Analyse …

… en composante principale [*principal component analysis*]
Méthode d'analyse multifactorielle* descriptive qui permet de déterminer les variables qui contribuent le plus à la variabilité observée, ainsi que les groupes de variables corrélées.

… discriminante [*discriminant analysis*]

Méthode d'analyse multifactorielle qui permet, à l'aide de covariables* de déterminer un score numérique permettant la discrimination optimale entre deux groupes de sujets A et B.

… factorielle de correspondance

Méthode d'analyse multifactorielle* descriptive applicable si les variables étudiées sont qualitatives. Son principe est assez proche de celui des analyses en composante principale*.

… intermédiaire [*intermediate analysis*]

Dans un essai randomisé*, analyse réalisée avant la fin de l'essai. Nécessite de réfléchir aux tests répétés et au contrôle du risque de première espèce.

… multifactorielle [*multivariate analysis*]

Sélectionne les covariables indépendantes entre elles et liées à la variable que l'on cherche à expliquer. Elles reposent sur le principe d'ajustement*. Voir : régression multiple*, régression logistique*, modèle de Cox*, analyse discriminante*.

… séquentielle [*sequential analysis*]

Il s'agit d'une procédure permettant les analyses intermédiaires. Consiste à faire une analyse cumulée après chaque événement ou après un groupe d'événements.

… sous-groupes

Analyse réalisée sur une partie d'un échantillon. Nécessite de corriger pour des tests multiples pour contrôler le risque de première espèce.

… unifactorielle [*univariate analysis*]

Étude des liens entre une variable expliquante (ou covariable) et une variable que l'on cherche à expliquer.

… de variance [*ANalyis Of Variance* ANOVA]

Compare les moyennes de plus de deux groupes dans des échantillons indépendants (ANOVA à un facteur) ou appariés* (ANOVA à deux facteurs). C'est une généralisation du test t de Student à plus de deux groupes.

Apparié (échantillons…)

Qualifie deux échantillons dans lesquels chaque observation correspond préférentiellement à une de l'autre échantillon.

Par exemple pression artérielle diastolique et systolique ou encore cholestérolémie avant et après traitement.

Association (force d'…)
Mesure l'intensité des liens qui peuvent exister entre deux variables. Voir : coefficient de corrélation*, risques relatif*, rapports de cote [*odds-ratio*]*.

Aveugle (prescription en…) [*blind*]
Dans un essai randomisé*, ignorance par le prescripteur de ce que le patient reçoit. Si patient et médecin sont dans cette ignorance (seul un tiers a cette connaissance) on parle de double aveugle ou de double « insu ».

B

Bayes
Pasteur britannique (XVIIIe siècle), auteur d'un théorème qui permet d'inverser le conditionnement dans une probabilité conditionnelle. Il permet par exemple de relier les valeurs prédictives d'un test à la sensibilité et spécificité de celui-ci. Voir : sensibilité, spécificité, valeurs prédictives.

beta, β (risque) [*beta type of error*]
Désigne le risque de deuxième espèce*.

Bilatéral (test) *[two tailed or two sided analysis*]
Se dit d'un test lorsque l'hypothèse alternative ne privilégie pas une direction pour la différence, c'est-à-dire ne teste pas spécifiquement une augmentation ou une diminution, mais l'une ou bien l'autre de ces possibilités. C'est la forme privilégiée des tests.
Unilatéral*.

C

Cas-témoins (enquête) [*case control study*]
Protocole emblématique de l'étude épidémiologique. Les participants sont inclus sur la base de leur statut vis-à-vis de la maladie d'intérêt, et non pas sur l'exposition à un facteur de risque. On a donc un groupe de « cas », présentant la maladie ; un groupe de « témoins » qui ne l'a pas. L'analyse consiste à comparer la fréquence d'exposition entre ces deux groupes.

Causalité [*causality*]
Relation de cause à effet entre un événement et un autre. Une question importante en épidémiologie est de décider si une association observée est causale.

Censurée (variable) [*censored data*]

Qualifie la durée jusqu'à un événement d'intérêt, lorsque le suivi est interrompu avant la réalisation de l'événement.

Ex : la durée de survie est censurée si le patient n'est pas suivi jusqu'à son décès.

Chi carré (χ^2) [*Chi square*]

Test statistique semi-paramétrique pour tester l'existence d'une association entre des variables catégorielles, qualitatives.

Clause d'ignorance

Désigne le secret d'attribution dans un essai randomisé*.

Cochran (test de) [*Cochran Q test*]

Test statistique pour estimer l'association entre des variables qualitatives en ajustant sur une tierce variable*.

Coefficient de corrélation [*coefficient of correlation*]

Valeur permettant l'estimation du degré d'association entre deux variables quantitatives. Le coefficient va de -1 à 1, la valeur 0 correspondant à l'absence d'association.

Coefficient kappa (κ) [*Kappa coefficient*]

Coefficient permettant de mesurer la concordance en excès du hasard.

Coefficient de Mahalanobis [*Mahalanobis coefficient*]

Distance standardisée utilisée notamment dans les analyses discriminantes*.

Coefficient de régression partielle [*coefficient of partial regression*]

Mesure la relation mathématique entre deux variables.

Cohorte (étude de) [*cohort study*]

Étude longitudinale d'un échantillon de sujets les uns exposés à un risque, les autres non exposés.

Combinaison [*combination*]

Nombre de façons de répartir *k* succès parmi *n* tentatives.

Comparaison historique [*historical comparison*]

Désigne une étude où les échantillons comparés ont été sélectionnés à des moments différents dans le temps.

Composante principale (analyse en*)

Voir analyse.

Concordance [*concordance*]
Méthode d'appréciation de l'accord entre plusieurs observateurs concernant un même patient.

CONSORT [*Consolidated standard of reporting trials*]
Ensemble de recommandations pour l'écriture d'un article détaillant les résultats d'un essai randomisé.

Continue (variable) [*continuous variable*]
Se dit de variables quantitatives prenant des valeurs réelles.

Corrélation (coefficient de)
Coefficient de corrélation*.

Corrélation partielle [*partial correlation*]
Mesure d'association entre deux variables quantitatives ajustée sur des variables tierces.

Courbe actuarielle [*Actuarial curve*]
Méthode actuarielle*.

Courbe de Kaplan Meier [*Kaplan Meier curve*]
Kaplan-Meier*.

Courbes de répartition
Pour variables quantitatives dont les valeurs sont portées en abscisse et les fréquences relatives cumulées en ordonnées.

Courbe ROC [*receiver operating characteristic*]
En ordonnées : sensibilité d'un examen ; en abscisse : 1 - la spécificité, estimées pour différentes valeurs seuil de l'examen.

Courbe de survie [*survival curves*]
Figuration graphique de variables censurées : actuarielles*, Kaplan Meier*.

Coût-bénéfice (étude) [*cost-benefit analysis*]
Relie les coûts à ses conséquences exprimées en unités monétaires.

Coût-efficacité (étude) [*cost-effectiveness analysis*]
Destinée à relier les coûts d'une action médicale à ses conséquences exprimées en unités physiques (critère d'efficacité).

Coût-utilité (étude) [*cost-usefulness analysis*]
Relie les coûts d'une action médicale à ses conséquences exprimées en variables qualitatives, nombre d'années de vie gagnées, années-qualité de vie, etc.

Covariable [*covariable*]
Désigne habituellement des variables « expliquantes » dans une analyse multifactorielle*.
Variable expliquante*.

Cox (modèle de) [*Cox model*]
Méthode d'analyse multifactorielle* pour des variables censurées*.

Critère de jugement
Résultat d'un examen, d'un traitement, d'un pronostic ou la survenue d'une maladie sur lequel est basée une comparaison entre échantillons. On distingue critère principal et secondaire.

D

Date des dernières nouvelles [*date of last news*]
Dans l'étude de la survie, date de l'événement (décès, récidive, etc.) ou date à laquelle le sujet a été revu pour la dernière fois.

Date de point [*date of follow-up*]
Dans l'étude de la survie, date à laquelle on cesse le suivi. Elle correspond à la date de censure pour tous les sujets n'ayant pas fait l'événement d'intérêt.

Degré de liberté [*degree of freedom*]
Paramètre associé à certaines distributions, comme le chi-carré, le Student. Le nombre de degré de liberté permet de choisir le seuil de rejet de l'hypothèse nulle dans les tests correspondants.

Delphi
La technique Delphi est une méthode de détermination de consensus à partir de jugements répétés d'un groupe d'experts.

Dépendante (variable) [*dependent variable*]
Variable dont la valeur se modifie avec les modifications d'une ou d'autres variables considérées dans l'étude.

Descriptives (méthodes multifactorielles) [*multivariate descriptive methods*]
Ces analyses, en composante principale, factorielle de correspondance situent sur un plan les covariables standardisées.

Deuxième espèce (risque) [*beta type of error*]
Risque beta*.

Discontinue (variable) [*discontinuous variable*]
Se dit de variables quantitatives prenant des valeurs entières.

Discrète (variable)
Variable discontinue*.

Discrétisation [*discretisation*]
Transformation d'une variable continue en une variable discrète par arrondi.

Discriminante (analyse) [*discriminant analysis*]
Analyse discriminante*.

Double aveugle [*double blind*]
Voir Aveugle*.

E

Écart-type [*standard deviation*]
Racine carrée de la variance. S'il s'agit de la valeur théorique (de la population) on la note s, s'il s'agit d'une valeur estimée à partir d'un échantillon, on note *s*.

Échantillon apparié
Voir Apparié*.

Échelle visuelle analogique [*visual analogic scale*]
Moyen de mesure de critères subjectifs de jugement.

Effectifs (d'une étude) [*numbe*r]
dans un essai randomisé, la détermination des effectifs est nécessaire pour limiter le risque de deuxième espèce.

Épidémiologie [*epidemiology*]
Étymologiquement, étude des épidémies des maladies transmissibles. Aujourd'hui, désigne l'étude des maladies, de leurs facteurs de risque, et des interventions d'un point de vue populationnel.

Épidémiologie analytique
Étudie les facteurs susceptibles de favoriser la survenue de maladies. Synonymes : épidémiologie explicative et prédictive.

Épidémiologie descriptive
Synonyme de connaissance des indicateurs de santé.

Épidémiologie explicative et prédictive
Epidémiologie analytique*.

Enquête (ou étude) cas-témoins [*case control study*]
Cas-témoins (enquête)*.

Enquête (ou étude) longitudinale [*cohort study*]
Cohorte*.

Enquête (ou étude) transversale [*cross-sectional study*]
Étude épidémiologique descriptive réalisée sur un échantillon donné, à un moment donné.

Équivalence
L'équivalence est obtenue lorsque la non-infériorité* est montrée dans les deux sens. L'équivalence n'est pas synonyme de test non significatif.

Erreur de première espèce [*alpha type of error*]
Risque de première espèce*.

Erreur de deuxième espèce [*beta type of error*]
Risque de deuxième espèce*.

Erreur de troisième espèce [*gamma type of error*]
Risque de troisième espèce. Très peu usité.

Essai croisé [*crossover trial*]
Administration à un sous-groupe d'un traitement A, puis d'un traitement B et à un autre sous-groupe du traitement B, puis A.

Essai randomisé [*randomized study*]
Consiste, au sein d'une population, à constituer deux (ou plus) sous-groupes par tirage au sort pour comparer entre eux deux (ou plus) examens complémentaires ou deux (ou plus) traitements qu'ils soient médicaux, chirurgicaux ou un traitement médical et un traitement chirurgical.

Éthique (d'un essai randomisé) [*ethical considerations for a randomized study*]
Il n'est licite d'entreprendre un essai randomisé qu'à la double condition presque paradoxale : à la fois espérer qu'un traitement est plus efficace qu'un autre (ou qu'un placebo) et douter de cette hypothèse.

Étude

coût-bénéfice
Relie les coûts à ses conséquences exprimées en unité monétaire.

coût-efficacité
Relie les coûts d'un traitement à ses résultats en termes de santé exprimés en unités physiques.

coût-utilité

Relie les coûts d'une action médicale à ses conséquences en termes médicaux.

multicentrique [*multicentric study*]

Étude menée en commun par plusieurs centres ou plusieurs équipes.

multifactorielle [multivariate study]

Estime les liens entre des variables expliquantes et une variable que l'on cherche à expliquer.

(ou enquête) transversale * [*cross-sectional study*]

multifactorielle [*univariate study*]

Estime les liens entre une variable expliquante et une variable que l'on cherche à expliquer.

(ou enquête) longitudinale * [*cohort study*]

Exclus-vivants

Dans l'estimation de la survie, les exclus-vivants sont les sujets qui ne sont pas décédés au moment de la date de point de l'étude, celle où l'on cesse de recueillir les nouvelles.

Expliquante (variable) [*covariable*]

Covariable*.

Expliquée (variable)

Variable que l'on cherche à expliquer, notamment dans une régression.

Exposés

En épidémiologie dans une enquête prospective, sujets exposés à un facteur de risque présumé.

F

Facteurs de confusion

Facteur responsable de la liaison observée entre deux autres variables. Un facteur de confusion peut mener à une conclusion erronée dans une étude épidémiologique.

Factorielle

Résultat de la multiplication de tous les nombres entiers inférieurs ou égaux à ce nombre (en excluant le zéro). On utilise le symbole « ! » pour noter cette opération.

Fisher (test exact de) [*Fisher exact test*]
Test statistique non paramétrique pour estimer les liens entre des variables qualitatives dans des échantillons ou groupes indépendants. Le test du chi-carré est une très bonne approximation du test de Fisher lorsque les échantillons sont grands.

Fonction de répartition
Synonyme : fréquences relatives cumulées.

Fréquence cumulée
Synonyme : fonction de répartition.

Friedman (test de) [*Friedman test*]
Test statistique non paramétrique pour comparer les distributions de plusieurs échantillons appariés.

G

Gauss (loi) [*Gaussian distribution*]
Loi normale*.

Gold standard [*gold standard*]
Stricto sensu étalon or. Anglicisme passé dans le langage courant, qui désigne le test de référence, celui qui permet de déterminer la présence ou non de la maladie.

Grades de recommandation [*degrees of recommandation*]
Niveaux de preuve scientifique. Ils sont exprimés par des lettres, le meilleur niveau correspondant à la lettre A.

H

Historique (comparaison) [*historical comparison*]
Comparaison historique*.

Hypothèse nulle [*null hypothesis*]
Dans un test statistique, l'hypothèse qui correspond au « statu quo », c'est-à-dire l'absence de différence, de corrélation, etc.

I

Ignorance (clause d') [*clause of ignorance*]
Dans un essai randomisé, c'est le fait qu'au moment de l'inclusion du sujet dans l'essai, le prescripteur ignore le traitement qui sera alloué au sujet.

Incidence [*incidence*]
En épidémiologie, nombre de nouveaux cas d'une maladie, recensés pendant une période de temps donnée (en général annuelle).

Insu [*blind*]
Aveugle*.

Intention de traiter (analyse en) [*intention to treat analysis*]
Au terme d'un essai randomisé, analyse des résultats selon le traitement qui a été théoriquement alloué par le tirage au sort, de tous les sujets randomisés.

Interaction [*interaction*]
On dit qu'il y a interaction entre deux facteurs de risque lorsque le risque relatif associé à la présence conjointe de ces deux facteurs diffère de celui qui serait conféré indépendamment par chacun des facteurs de risque.

Intermédiaire (analyse) [*intermediate analysis*]
Analyse intermédiaire*.

Intervalle de confiance [*confidence interval*]
Estimation, à partir d'un pourcentage observé sur un échantillon, de la fourchette dans laquelle aurait 95 % de chances de se situer la réalité.

K

Kaplan-Meier (méthode) [*Kaplan-Meier estimation*]
Méthode non paramétrique d'estimation de la probabilité de survie. L'estimation repose sur le principe des probabilités conditionnelles. La probabilité de survie est recalculée après chaque événement, en prenant en compte le nombre de personnes censurées. C'est la méthode d'étude de la survie la plus utilisée en épidémiologie clinique.

Kappa (coefficient) [*Kappa value*]
Coefficient qui permet de quantifier la concordance entre deux mesures.

Kappa pondéré (coefficent)
Coefficient qui permet de quantifier la concordance entre deux mesures, en donnant moins de poids aux discordances légères.

Kruskall-Walis (test de) [*Kruskall-Walis test*]
Test statistique non paramétrique pour comparer la distribution d'une variable quantitative entre plus de deux échantillons. Utilisée dans les

mêmes circonstances que l'analyse de Variance, extension du test de Wilcoxon-Mann-Whitney à plus de deux échantillons.

L

Laplace-Gauss (loi de) [*Gaussian distribution*]
Loi normale*.

Loi binomiale [*binomial distribution*]
Loi de distribution de variables discontinues lorsqu'elle s'applique à des données qui ont des caractéristiques binaires.

Loi normale [*Gaussian distribution*]
Loi de distribution de variables quantitatives continues définie par la moyenne et la variance

Loi de Poisson [*Poisson distribution*]
Loi de distribution de variables discontinues. Dans le cas d'événements rares, c'est une bonne approximation de la loi binomiale.

Logistique (régression) [*logistic regression*]
Régression logistique*.

Logrank (test du) [*Logrank test*]
Test statistique paramétrique pour estimer les liens entre des variables qualitatives et des variables censurées.

Longitudinale (étude) [*cohort study*]
Cohorte*.

M

McNemar (test de) [*McNemar test*]
Test statistique paramétrique pour comparer le pourcentage de succès lorsque les échantillons sont appariés.

Mahalalobis (coefficient de) [*Mahalanobis coefficient*]
Coefficient de Mahalanobis*.

Mann-Whitney (test de) [*Mann-Whitney test*]
Test statistique non paramétrique pour comparer la distribution d'une variable quantitative entre deux échantillons. S'utilise dans les mêmes circonstances que le test de Student*. Aussi connu comme test de Wilcoxon-Mann-Whitney.

Mantel-Haenszel (test de) [*Mantel-Haenszel test*]
Test statistique non paramétrique pour estimer les liens entre plusieurs variables qualitatives et une variable qualitative. Aussi connu comme test de Cochran Mantel Haenszel.

Médiane [*median*]
Valeur qui partage l'échantillon ordonné en deux parties d'effectif égales.

Méta-analyse [*meta-analysis*]
Étude synthétisant toutes les données recueillies sur le sujet et se différenciant d'une simple synthèse bibliographique.

Méthode actuarielle [*actuarial method*]
Actuarielle*.

Méthodes descriptives [*descriptive methods*]
Types d'analyses multifactorielles comme l'analyse en composante principale ou l'analyse factorielle de correspondance.

Méthode de Kaplan-Meier [*Kaplan-Meier estimation*]
Kaplan-Meier*.

Méthodes prédictives [*predictive methods*]
Types d'analyses multifactorielles.

Modèle de Cox.
Cox*.

Moyenne arithmétique [*mean*]
Somme des valeurs observées divisée par le nombre de variables observées.

Multicentrique (étude) [*multicentric study*]
Étude multicentrique*.

N

Nombres au hasard [*hazard numbers*]
Séries de nombres obtenus par une procédure aléatoire.

Nominale (variable)
Variable qualitative. Elle peut être à deux ou plusieurs classes et, dans ce dernier cas être ordonnée ou non.

Non-inferiorité

La non-fériorité d'un traitement est montrée lorsque l'efficacité n'est pas inférieure de plus d'une marge, fixée *a priori*, de l'efficacité du traitement de référence.

Normale (loi) [*Gaussian distribution*]

Loi normale*.

O

Odd n.m.

Terme anglais signifiant « cote ».

Odds-ratio

Anglicisme signifiant « rapports de cotes ».

Ordonnée (variable qualitative)

Variable qualitative dont les valeurs ont un ordre logique ; par exemple « Absent »/« Faible »/« Moyen» /« Fort » pour l'expression d'un symptôme.

P

P (ou P value)

Probabilité qu'une différence égale à la différence observée avait, d'être obtenue s'il n'y avait pas de différence entre les interventions comparées. Egalement appelé degré de signification.

Participation à une étude (temps de) [*contribution period*]

Délai entre la date des dernières nouvelles et la date d'origine.

Pas à pas (analyse) [*stepwise analysis*]

Méthode d'analyse multifactorielle qui consiste à juger séquentiellement l'introduction ou le retrait des variables explicatives.

Pearson (coefficient de corrélation de) [*Pearson correlation coefficient*]

Test statistique paramétrique pour estimer l'association entre deux variables quantitatives.

Permutation de nombres aléatoires (table de) [*permutation random number table*]

Tables utilisées pour répartir de manière aléatoire des sujets inclus dans un essai randomisé afin de produire un tirage au sort équitable.

Perdus de vue (sujets) [*lost of follow-up*]
Sujets en vie lors de la date des dernières nouvelles si elle est antérieure à la date de point.

Pertinence clinique [*clinical relevance*]
Permet de s'assurer que le résultat d'un essai randomisé a un effet suffisamment important.

Peto (méthode de) [*Peto method*]
Méthode appliquée dans les méta-analyses pour calculer les rapports de cotes selon un modèle à effet fixe, c'est-à-dire en l'absence d'hétérogénéité.

PICO acronyme [pour *Patient Intervention Control Outcome*].

Placebo (effet) [*placebo effect*]
Effet, souvent positif, psychologique, physiologique ou psychophysiologique de tout produit non lié au principe actif.

Point (date de) [*date of follow-up*]
Date de point*.

Poisson (loi de) [*Poisson distribution*]
Loi de Poisson*.

Pondération [*weighting*]
Attribution à chacun des éléments servant à élaborer une moyenne ou un indice ou un score, d'un coefficient qui exprime son importance relative.

Précision
En statistique, valeur correspondant à la demi-largeur de l'intervalle de confiance.

Précision diagnostique [*diagnostic accuracy*]
Estimation de la valeur globale d'un test diagnostique.

Prédictives (méthodes multifactorielles) [*multivariate analysis*]
Multifactorielle (analyse)*.

Prédictives (valeurs) [*predictive values*]
Probabilités au vu du résultat d'un test que le sujet ait ou n'ait pas la maladie expliquée par ce test.

Première espèce (risque de) [*alpha risk*]
Risque de première espèce*.

Prévalence [*prevalence*]
Dans une population, nombre de cas (anciens et nouveaux) observés à un instant donné.

Probabilités bayésiennes [*bayesian probabilities*]
Bayes*.

Probabilités conditionnelles [*conditional probabilities*]
Probabilité d'un événement si un autre événement est présent.

Propension (score)
Score de propension*.

Puissance (d'un test) [*test powerfull*]
Dans une comparaison, complément du risque de deuxième espèce (1 – β).

Q

Qualité de vie [*quality of life*]
La qualité de vie appliquée à la santé (*Health Related Quality of Life*) est estimée par des scores généraux ou plus ou moins spécifiques d'une maladie ou d'un ensemble de maladies.

QUOROM acronyme [pour *quality of reporting of meta-analyses*]
Liste d'items à laquelle doit satisfaire une méta-analyse d'essais randomisés.

R

Randomisation [*randomisation*]
Essai randomisé*.

Rangs [*ranks*]
Les tests non paramétriques reposent sur la notion de rangs et s'affranchissent ainsi de la contrainte de distribution normale qui est exigée pour utiliser des tests paramétriques.

Rapports de hasards [*hazards ratio, HR*]
Risque relatif de survenue d'un événement dans une analyse multifactorielle réalisée selon le modèle de Cox.

Rapports de cotes [*odds ratio*]
Cette mesure approche de façon correcte le risque relatif dont l'utilisation n'est motivée que pour des raisons mathématiques.

Rapport de vraisemblance [*likelihood ratio*]
Dans un examen complémentaire, rapport du pourcentage de « vrais positifs » chez les malades à celui du pourcentage de « faux positifs » chez les sujets inclus dans l'étude, mais qui n'ont pas la maladie.

Référentiel [*frame of reference*]
Critère externe de jugement de la présence ou non d'un signe ou d'une maladie.

Régression [*regression*]

linéaire [*linear regression*]
Étudie et mesure la relation mathématique qui peut exister entre deux (régression simple) ou plusieurs (régression multiple) variables, lorsque la variable dépendante est quantitative.

logistique [*logistic regression*]
Méthode d'analyse multifactorielle utilisable lorsque la variable expliquée est qualitative à deux classes, les covariables étudiées pouvant être quantitatives ou qualitatives.

multiple [*multiple linear regression*]
Régression intégrant plusieurs variables explicatives.

partielle (coefficient de) [*coefficient of partial regression*]
Coefficient de régression partielle*.

simple [*single linear regression*]
Régression linéaire*.

REVMAN® [*Review Manager*]
Logiciel développé par la Cochrona collaboration comme aide à la recherche de revue systématique ou en vue d'une méta-analyse. « www.cc-ims.net/RevMan/download.htm. »

Risque [*risk*]

absolu [*absolute risk*]
Mesure de fréquence du nombre de nouveaux cas sur l'effectif de la population étudiée pendant une période donnée (incidence) ou de cas existants à un instant donné (prévalence).

attribuable [*attributable risk*]
Fraction étiologique d'un risque pouvant être soit celui chez les sujets exposés au risque, soit et le risque attribuable en population.

de première espèce [*alpha type of error*]
Alpha*.

de deuxième espèce [*beta type of error*]
Beta*.

de troisième espèce [*gamma type of error*]
Dans une comparaison, risque de conclure à tort qu'un élément de la comparaison est supérieur à un autre alors que c'est l'inverse.

en excès
Différence entre le risque de survenue d'une maladie chez les sujets exposés au risque et le risque de cette même maladie chez les sujets non exposés.

relatif [*relative risk*]
Rapport entre le risque chez les sujets exposés au risque et le risque chez les sujets non exposés.

Risques proportionnels (modèle de) [*proportional hasard model*]
Hypothèse nécessaire à l'utilisation du modèle de Cox pour les données censurées. Cox*.

ROC (courbe) [*receiving operative characteristic curves*].
Courbe ROC*.

S

Séquentielle (analyse) [*sequential analysis*]
Analyse séquentielle*.

Sensibilité [*sensitivity*]
Probabilité qu'un signe diagnostique soit présent chez un individu atteint d'une maladie.

Score pronostique [*prognosis score*]
Estimation d'une probabilité pronostique, basée sur des coefficients calculés à partir des risques relatifs d'une étude multifactorielle.

Score de propension [*propensity score*]
Il désigne la probabilité qu'avait une personne de recevoir une intervention dans une étude observationnelle.

Score de qualité de vie [*score of life quality*]
Qualité de vie*.

Secret d'attribution [*concealment of allocation*]
Dans un essai randomisé, au moment ou un sujet est inclus dans l'essai, personne ne doit savoir *a priori* du traitement qui lui sera alloué.

Spearman (coefficient de corrélation) [*Spearman coefficient*]
Test statistique non paramétrique pour estimer l'association entre deux variables quantitatives.

Spécificité [*specificity*]
Probabilité qu'un signe diagnostique soit absent chez un individu non atteint d'une maladie.

Standard de référence externe [*frame of reference*]
Référentiel*.

Standardisation [*standardizatio*n]
Action de rapporter une mesure à son écart-type.

STARD [*Standard for reporting of diagnostic accuracy*]
Liste d'items pour améliorer la qualité des publications sur les moyens diagnostiques.

Statistiquement significative (différence) [*statistically significant*]
Dans une comparaison, lorsque la différence observée est jugée incompatible avec le seul effet du hasard.

Stratification [*stratification*]
Dans un essai randomisé, si un facteur de pronostic est important, au lieu de faire un tirage au sort sur l'ensemble de cette population, on fait un tirage au sort parmi les malades qui ont ce facteur de pronostic et un autre chez ceux qui ne l'ont pas. On dit dans ce cas que la randomisation est stratifiée sur ce facteur.

STROBE [*Strengthening the reporting observational studies in epidemiology*]
Grille d'élaboration et d'évaluation des études observationnelles.

Student (test de) [*Student test*]
Test statistique paramétrique pour comparer la moyenne d'une variable quantitative entre deux échantillons.

Supériorité
La démonstration de supériorité d'un traitement sur un autre est obtenue en cas de test statistique significatif.

Survie actuarielle [*actuarial survival*]
Actuarielle (méthode)*.

Survie cumulée [*cumulative survival*]
Taux de survie à un temps *t* qui est le produit des taux de survie antérieurs par le taux de survie dans le dernier intervalle.

T

***t* (test)** [*t test*]
Test de Student*.

Taux d'incidence.
Incidence*.

Test bilatéral [*bilateral test*]
Bilatéral*.

Test paramétriques [*parametric tests*]
Tests utilisés lorsque les variables étudiées suivent une distribution que l'on peut décrire mathématiquement à partir de paramètres comme la loi normale.

Tests non paramétriques [*non parametrics tests*]
Tests qui ne demandent pas d'hypothèse sur la distribution des variables étudiées.

Tirage au sort [*randomisation*]
Base des essais randomisés assurant les meilleures chances que les groupes comparés soient similaires.

Transversale (étude)
Epidémiologie descriptive*.

Tronquée (variable) [*censored data*]
Censurée (variable)*.

U

Unilatéral (test) [*unilateral test*]
Se dit d'un test lorsque l'hypothèse alternative privilégie une direction pour la différence d'intérêt : augmentation ou diminution. Bilatéral*.

V

Valeurs prédictives [*predictive values*]
Prédictives (valeurs)*.

Variable [variable]

aléatoire [*random variable*]
Variable dont les mesures sont soumises à des fluctuations d'échantillonnage.

appariée [*match variable*]
Apparié*.

censurée (ou tronquée) [*censored variable*]
Censurée (variable)*.

continue [*continuous variable*]
Variables quantitatives qui prennent des valeurs réelles avec de nombreuses décimales.

dépendante
Variables expliquée dans une régression.

discrète [*discrete variable*]
Variables quantitatives qui prennent des valeurs entières.

expliquante [*explainatory variable*]
Variable susceptible d'influencer une autre variable que l'on cherche à expliquer dans une régression.
Covariable*.

expliquée (ou dépendante) [*explained variable*]
Variable que l'on cherche à expliquer.

qualitative (ou nominale ou catégorielle) [*nominal variable*]
Variables qui peuvent prendre des valeurs distinctes.

quantitative
Variable numérique.

discrète [*discrete variable*]
Variable discontinue*.

nominale [*nominal variable*]
Variable qualitative*.

tronquée [*censored variable*]
Censurée (variable)*.

Variance
Valeur quantifiant l'étendue de la dispersion des valeurs autour de la moyenne.

Vraisemblance (rapport de) [*likelihood ratio*]
Rapport de vraisemblance*.

W

Wilcoxon (test) [*Wilcoxon rank sum test*]
Test statistique non paramétrique pour comparer la distribution d'une variable quantitative entre deux échantillons. Mann-Whitney*.

X

X² (ou plutôt χ^2) [*Chi square*]
Chi carré*.

Y

Yates (correction de) [*Yates correction*]
Correction parfois employée dans le test du chi-carré. On préfère cependant utiliser le test exact de Fisher. Fisher*. Chi carré*.

Formaté typographiquement par DESK (53) :
02 43 01 22 11 – desk@desk53.com.fr

Impression & brochage sepec - France
Numéro d'impression : 04495130722 - Dépôt légal : août 2013



PEFC 10-31-1470 / Certifié PEFC / Ce produit est issu de forêts gérées durablement et de sources contrôlées. / pefc-france.org